DOCUMENT RESUME

ED 426 105 TM 029 335

AUTHOR	Herman, Joan L.; Aschbacher, Pamela R.; Winters, Lynn
TITLE	Guia Practica para una Evaluacion Alternativa (Practical Guide to Alternative Assessment).
INSTITUTION	Association for Supervision and Curriculum Development, Alexandria, VA.; Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
ISBN	ISBN-0-87120-285-9
PUB DATE	1997-00-00
NOTE	131p.; Translation of "A Practical Guide to Alternative Assessment"; translated by Maria A. Brito Medina and Gina Oxbrow. For the English version, see ED 352 389.
CONTRACT	R117G10027
AVAILABLE FROM	Association for Supervision and Curriculum Development, 1250 N. Pitt Street, Alexandria, VA 22314-1453 (Order number 896293; $12.95, members; $15.95, nonmembers).
PUB TYPE	Books (010) -- Guides - Non-Classroom (055)
LANGUAGE	Spanish
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	*Cognitive Processes; Curriculum Development; *Decision Making; *Educational Assessment; Educational Theories; Elementary Secondary Education; Guidelines; *Performance Based Assessment; Scoring; *Test Construction; Test Use; Trend Analysis
IDENTIFIERS	*Alternative Assessment; Process Models

ABSTRACT

Guidance is offered in Spanish on the creation and use of alternative assessment, and a process model is presented that links assessment with curriculum and instruction, based on contemporary theories of learning and cognition. The introductory chapter provides background on the purposes of assessment and the need for new alternatives, with an overview of key assessment development issues. Linking assessment and instruction is the focus of chapter 2, which also reviews current trends in assessment. Chapter 3 considers determining the purpose of the assessment, and chapter 4 reviews selecting assessment tasks and matching them to student outcomes. Setting the criteria for judging student performance is discussed in chapter 5. Chapter 6 reviews the steps necessary to ensure reliable scoring. Chapter 7 makes the important point that assessment is not an end in itself, but rather a tool for decision making. In this context, reliability and validity of assessments are discussed. (Contains 26 figures.) (SLD)

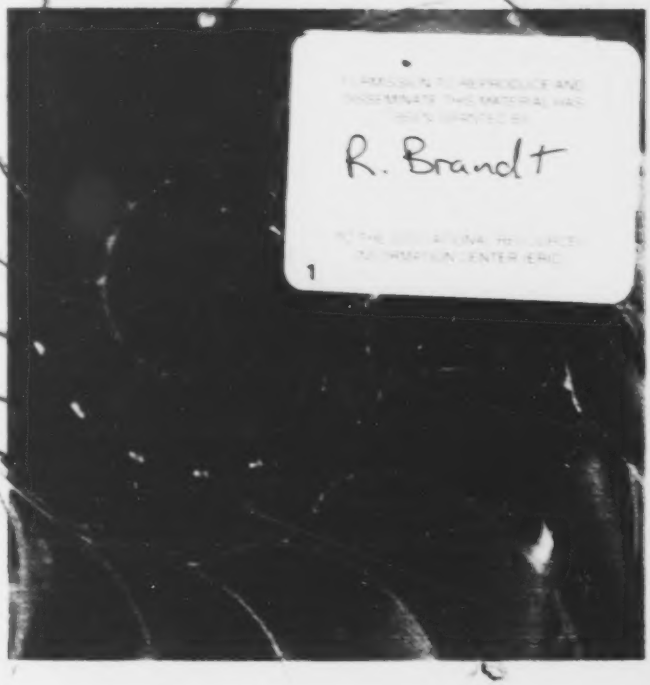
****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 426 105

Guía Práctica para una Evaluación Alternativa

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Joan L. Herman
Pamela R. Aschbacher
Lynn Winters



PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
R. Brandt
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

TM029335

BEST COPY AVAILABLE

Guía Práctica para una Evaluación Alternativa

Joan L. Herman
Pamela R. Aschbacher
Lynn Winters

ASCD

Association for Supervision and Curriculum Development

Alexandria, Virginia, USA

CRESST

National Center for Research on Evaluation, Standards, and Student Testing

University of California, Los Angeles
Los Angeles, California, USA

BEST COPY AVAILABLE

Association for Supervision and Curriculum Development
1250 N. Pitt Street• Alexandria, VA 22314-1453
Teléfono: 1-800-933-2723 703-549-9110 • Fax: 703-299-8631

Las publicaciones de ASCD contienen gran variedad de opiniones. Las opiniones expresadas o implícitas en esta publicación no deben interpretarse como la postura oficial de la Asociación.

Título original: *A Practical Guide to Alternative Assessment*
Traducción: María A. Brito Medina
Gina Oxbrow

Copyright © 1997 Association for Supervision and Curriculum Development.

Este volumen es una traducción fiel y exacta del original titulado *A Practical Guide to Alternative Assessment*. Este trabajo fue apoyado por el convenio R117G10027 del Educational Research and Development Center Program y con número de catálogo CFDA 84.117G administrado por la Office of Educational Research and Improvement, U.S. Department of Education. Las conclusiones y opiniones expresadas en este trabajo no reflejan la postura o política de la Office of Educational Research and Improvement o del U.S. Department of Education.

Impreso en los Estados Unidos de América.

Número de artículo de ASCD: 896293 s4/97
Precio para miembros de ASCD: US$12,95; precio para no miembros: US$15,95

Library of Congress Cataloging-in-Publication Data

Herman, Joan L.
[Practical guide to alternative assessment. Spanish]
Guía práctica para una evaluación alternativa / Joan L. Herman, Pamela R. Aschbacher, Lynn Winters ; [traducción, María A. Brito Medina, Gina Oxbrow].
p. cm.
Includes bibliographical references.
Contents: Reconsiderar la evaluación -- Unir la evaluación y la enseñanza -- Establecer los objetivos -- Seleccionar tareas de evaluación -- Establecer criterios -- Asegurar una puntuación justa -- Utilizar la evaluación alternativa en la toma de decisiones.
ISBN 0-87120-285-9 (pbk.)
1. Educational tests and measurements--United States.
I. Aschbacher, Pamela R. II. Winters, Lynn. III. Title.
LB3051.H4518 1997
371.27'0973--DC21 97-4643
CIP

Guía práctica para una evaluación alternativa

Prólogo v

1. Reconsiderar la evaluación 1
2. Vincular la evaluación y la enseñanza 12
3. Establecer los objetivos 23
4. Selección de tareas de evaluación 33
5. Establecer criterios 44
6. Asegurar una calificación fiable 80
7. Utilización de la evaluación alternativa para la toma de decisiones 95

Acerca de los autores 123

Prólogo

Se dice que un cartel con la frase "No todo lo que cuenta puede contarse, ni todo lo que puede contarse cuenta" estaba colgado en la pared del despacho de Albert Einstein. Dentro del contexto de la reciente polémica sobre las reformas educativas, esta afirmación casi profética repercute sobre la evaluación del aprendizaje de los alumnos.

La evaluación se ha convertido en el enfoque central de la agenda actual de las reformas educativas de nuestro país. Aunque nuestro diálogo sobre una evaluación auténtica se ha llevado más allá de lo que es la medición de demostraciones de rendimientos humanos complejos puramente cuantificables o "contables", nos ha faltado un marco exhaustivo, sistemático e integrado para ayudar a los profesionales en el diseño y desarrollo de métodos de evaluación alternativa.

En *Guía práctica para una evaluación alternativa*, Joan Herman, Pamela Aschbacher y Lynn Winters nos ofrecen consejos convincentes para crear y utilizar métodos alternativos que miden los logros del alumno. Nos presentan un modelo sistemático, integrado e iterativo que vincula la evaluación con el currículo y la docencia, y que está fundamentado en las últimas teorías de aprendizaje y cognición.

Las autoras analizan los fines de la evaluación y esgrimen un argumento sustancial respaldatorio de las estrategias alternativas que se proponen. Sin embargo, como ellas mismas apuntan, el tema fundamental del libro es destacar varios temas clave relacionados con la evaluación que reafirman nuestra idea de que hay que incorporar los elementos más importantes de la práctica docente a las tareas de evaluación. Entre estos temas se incluyen:

1. La evaluación debe ser congruente con los objetivos docentes más importantes.
2. La evaluación debe incluir un análisis tanto de los procesos como de los productos del aprendizaje.

3. Las actividades con base en el rendimiento no constituyen una evaluación en sí.
4. La teoría del aprendizaje cognitivo y su enfoque constructivista de la adquisición de conocimientos respaldan la necesidad de integrar las metodologías de la evaluación con los fines pedagógicos y el contenido curricular.
5. Una visión integrada y activa del aprendizaje del alumno requiere de una evaluación del rendimiento integral y complejo.
6. El diseño de una evaluación está íntimamente relacionado con la finalidad y objetivos de la evaluación; la calificación y el seguimiento del progreso de un alumno se deben considerar materia aparte del diagnóstico y el mejoramiento.
7. La clave de una evaluación eficaz es la adecuación de la tarea al resultado que se desea del alumno.
8. Los criterios que se emplean para evaluar el rendimiento del alumno son cruciales; en ausencia de criterios, la evaluación sería sólo una actividad aislada y episódica.
9. Una buena evaluación brinda un gran número de datos que permite tomar decisiones con conocimiento de causa sobre el aprendizaje del alumno.
10. Los sistemas de evaluación que proporcionan la retroalimentación más exhaustiva sobre el progreso del alumno incluyen numerosas medidas que se han tomado con el tiempo.

La palabra "*assess*" (evaluación) proviene de la palabra francesa "assidere" que significa "sentarse al lado de". Las autoras, al clarificar cuáles son los aspectos conceptuales y técnicos cruciales en la utilización de evaluaciones alternativas, han reafirmado el papel fundamental que juega la evaluación, que es la provisión de retroalimentación auténtica y significativa para mejorar el aprendizaje del alumno, la calidad docente y las opciones educativas.

Como afirman las autoras, la evaluación no es simplemente un fin en sí misma. Es un proceso que facilita tomar decisiones pedagógicas acertadas al proporcionar información sobre dos preguntas fundamentales: ¿qué tal vamos? y ¿cómo podemos mejorar todo el proceso?

Quizás la mejor forma de responder a estas preguntas sería sentándonos al lado del alumno y tratar de averiguarlo. ¡Qué alternativa más interesante!

STEPHANIE PACE MARSHALL
Presidenta de la ASCD, 1992–93

1

Reconsiderar la evaluación

La evaluación es la piedra angular de la reforma educativa de la década de los años noventa: la agenda educativa del Presidente, América 2000; los Objetivos Nacionales de Educación establecidos por los gobernadores; el interés por ser competitivo a nivel internacional; las nuevas demandas para la reestructuración y la llamada *accountability*–responsabilidad adjudicada o rendir cuentas–a nivel estatal, local y escolar. Estas enérgicas y visibles iniciativas invitan a los educadores y a la nación a centrarse en objetivos de alto nivel para nuestros hijos. Nos alientan a ir en pos de la perfección y a dirigir nuestros esfuerzos hacia su consecución por el bien de los alumnos, escuelas, distritos, estados y de la nación. Al pedir que evaluemos el progreso, con frecuencia se presenta la evaluación como una llave para conseguir tal progreso y así se asegura el carácter prioritario de la evaluación en las escuelas.

Sin embargo, este mayor énfasis en la evaluación surge en una época de creciente descontento con los métodos tradicionales de exámenes tipo test (exámenes donde se presentan respuestas preseleccionadas). Como resultado se ha desatado un gran interés por los métodos alternativos de evaluación y se están llevando a cabo proyectos en todo el país para tratar de formularlos tanto a nivel nacional, estatal y local como en las aulas mismas. Los temas sobre carpetas de trabajos, exposiciones, experimentos prácticos y la expresión escrita en todo el currículo se han expuesto un sinnúmero de veces. A pesar de los múltiples congresos y reuniones en los que se han tratado estos temas, los educadores siguen

sin tener esa ayuda concreta que les permita formular y utilizar métodos alternativos de evaluación.

El propósito de este libro es contribuir al proceso de forjar métodos alternativos de evaluación. Está dirigido a maestros en formación y a los que ya ejercen la profesión, al personal directivo de escuelas y a profesionales a nivel de distrito y estatal que están interesados en nuevas formas de evaluación. Basado en teorías recientes sobre aprendizaje significativo y del currículo, así como en criterios de calidad de evaluación que ya han sido establecidos y aquéllos todavía en desarrollo, en este libro se propone un método sistemático para la elaboración de la evaluación y se plantean puntos críticos para asegurar evaluaciones de alta calidad. En este primer capítulo, analizamos tanto los fines de la evaluación como la necesidad de buscar nuevas alternativas; asimismo resumimos los temas clave en el desarrollo de la evaluación, lo que constituye el eje central de este libro.

También es importante señalar lo que este libro no pretende ser. No intenta ser un manual sobre cómo planificar y poner en práctica un sistema de evaluación exhaustivo o cómo elaborar un programa de evaluación para toda una clase. Por el contrario, su propósito es destacar los asuntos clave en el desarrollo de una única evaluación eficaz, componente importante en la realización de evaluaciones de calidad.

Definición de términos

Cuando se habla de las diversas alternativas para los exámenes tradicionales tipo test, se ventilan muchos términos. Entre éstos, tenemos evaluación alternativa, evaluación auténtica y evaluación basada en el rendimiento. Utilizamos estos términos como sinónimos de las variantes de la evaluación del rendimiento que exigen que el alumno genere una respuesta en lugar de escoger entre las respuestas que se ofrecen. La evaluación del rendimiento, nómbrese como se nombre, exige que los alumnos realicen de manera activa tareas complejas y significativas, a la vez que valora los conocimientos previos, el aprendizaje reciente y las destrezas necesarias para resolver problemas reales o auténticos. Algunos de los métodos alternativos de evaluación que nos vienen a la mente al emplear el término "evaluación alternativa" son exposiciones, investigaciones, demostraciones, respuestas orales y escritas, diarios y carpetas de trabajos.

Entender lo que la evaluación promete

¿Por qué se presta tanta atención a los exámenes o a otras formas de evaluación? ¿Por qué los necesitamos tanto? La evaluación cubre necesidades a todos los niveles de la jerarquía educativa. Por ejemplo, la evaluación ayuda a los educadores a establecer criterios, a crear objetivos para la docencia, a mejorar el rendimiento, a proporcionar retroalimentación diagnóstica, a calificar/evaluar el progreso y comunicarlo a otros.

Ya seamos maestros que utilizan exámenes tradicionales o coordinadores que preparan exámenes de rendimiento, los exámenes son el vehículo mediante el cual establecemos y comunicamos *criterios* a los que nos rodean: indicamos aquello que es importante, en lo que hay que centrarse y lo que representa un buen rendimiento. Durante este proceso los resultados de los exámenes están vinculados a intereses importantes—calificaciones finales, criterios de ingreso para la universidad, seguridad profesional, autosatisfacción y otras ventajas—y como consecuencia motivan el rendimiento. No sólo comunicamos a los alumnos lo que es importante cuando incluimos un tema en un examen, sino que también les motivamos a aprenderlo. Los coordinadores nacionales de exámenes recomiendan lo que se debe enfatizar en las escuelas y motivan, tanto a maestros como a alumnos, a que obtengan provecho de sus exámenes.

De igual manera, la retroalimentación y el seguimiento del progreso que forman parte de la evaluación funcionan a varios niveles. Para el personal directivo y los planificadores de las escuelas, los resultados de los exámenes proporcionan información sobre la eficacia de un programa e identifican los puntos fuertes y débiles de los currículos. Esto demuestra ser útil para la distribución de recursos, la identificación de necesidades de capacitación del personal docente o de diseño de materiales nuevos, y para la selección y evaluación de planes de mejora. Para los maestros, los exámenes proporcionan importante información diagnóstica que permite formar grupos por niveles, identificar las necesidades para la docencia y recomendar una enseñanza adecuada, determinar lo que se considera el dominio y evaluar la eficacia de unidades didácticas o métodos didácticos particulares. Para los padres y alumnos, la información procedente de exámenes es un indicador del progreso individual, que les permite entender y aprovechar los puntos fuertes y débiles del individuo.

Para todos, los exámenes prometen responder a las preguntas: *"¿Qué tal voy [vamos]? ¿Cómo podría [podríamos] mejorar?"*

Los exámenes sólo pueden cumplir esta promesa si reúnen ciertas condiciones indispensables. Entre éstas la más importante es el significado del rendimiento en un examen: los exámenes son útiles y productivos en tanto que representan metas *importantes* para los alumnos y objetivos valiosos para la enseñanza. En otras palabras, para que el contenido de un examen sea válido, justo y útil, éste debe ajustarse a los conocimientos, destrezas y disposiciones que enseñan los maestros y aquéllos que se espera que los alumnos aprendan o adquieran.

La ilustración 1.1 muestra un simple modelo que destaca la manera de utilizar sistemáticamente los resultados de la evaluación para respaldar y facilitar la mejora de la calidad docente. Como muestra la ilustración, las escuelas y los maestros generalmente sintetizan los datos procedentes de varias fuentes con el fin de llegar a los objetivos para los alumnos a nivel de escuela o de aula. Entre estas fuentes se encuentran las expectativas de la sociedad, las directrices curriculares estatales y de distrito, los requisitos legales, los textos y otros materiales didácticos disponibles, y los criterios y juicios profesionales. Una vez definidos, estos objetivos o metas sirven como señales indicadoras para la programación de la enseñanza y de la evaluación. Puesto que reflejan los mismos objetivos que gobiernan las actividades didácticas, los resultados de una evaluación guían la

Ilustración 1.1
Un modelo integrado

planificación de la enseñanza y sirven como medida de la calidad docente. Los resultados de la evaluación pueden utilizarse para identificar aquellas áreas donde determinados individuos necesitan más ayuda, o donde se necesita una enseñanza suplementaria, o donde se pueden mejorar las unidades didácticas, o donde hay que dirigir más esfuerzos a los recursos de desarrollo del personal, etcétera. Cuando la enseñanza y la evaluación se unen a una serie de objetivos de aprendizaje *significativos*, las evaluaciones tienen sentido y pueden usarse para mejorar la enseñanza.

No es que los exámenes deban regir el currículo, o que los maestros deban enseñar pensando exclusivamente en el examen. Más bien, *una buena evaluación*[1] es *un componente integral de una buena enseñanza*. Tanto los exámenes como la enseñanza deben reflejar metas significativas y preestablecidas que los alumnos han de alcanzar. Las evaluaciones deben medir los objetivos importantes del aula: los resultados de la evaluación deben concordar con el rendimiento de los alumnos en aquellas áreas de conocimientos generales y destrezas reflejadas en esos objetivos; y la enseñanza debe brindar a los alumnos la posibilidad de aprender y adquirir los conocimientos y destrezas.

Entender los límites de la evaluación tradicional

De las recientes críticas han surgido preguntas sobre la relación que existe entre el modelo que se muestra en la ilustración 1.1 y los ejercicios de evaluación existentes. ¿Son las notas de los exámenes un fiel reflejo de un proceso significativo de aprendizaje? ¿Representa una mejoría en las notas un mejor aprendizaje (Cannell 1987, Linn et al. 1990, Shepard 1989)? ¿Cómo es posible que casi todos los estados afirmen tener notas "superiores a la media" cuando las comparan con el modelo representativo nacional? La idea de "media" comparada con el modelo representativo nacional sugiere que algunos obtienen notas inferiores, otros superiores y otros al nivel de la media. ¿Se deben las mejoras de las notas de exámenes a una mejora de la enseñanza y del aprendizaje, o reflejan un currículo deficiente que implica que los alumnos están siendo "adiestrados y aniquilados" siguiendo el contenido previsto de los exámenes?

La interminable letanía sigue. Muchos se preguntan si los exámenes estándar actuales son suficientemente representativos de los importantes objetivos del aprendizaje y del desarrollo del alumno. Entre las críticas se incluyen el cerrado contenido de los exámenes que se concentra principalmente en las destrezas básicas de comprensión de lectura, lengua y matemáticas; la falta de correspondencia entre el contenido de los exámenes y el currículo y la enseñanza; el excesivo énfasis en las destrezas discretas y comunes que deja a un lado las de razonar y resolver problemas; y la poca relevancia de los exámenes tipo test en el

[1]Aunque a lo largo del libro venimos utilizando examinar y evaluar como más o menos sinónimos, preferimos el término evaluar porque nos lleva a pensar más allá de las definiciones tradicionales de lo que es un examen.

aprendizaje del aula o del mundo real (Baker 1989, Shepard 1989, Herman y Golan 1990). ¿Pueden llegar a conseguir resultados significativos los programas educativos que se basan en los tradicionales exámenes estándar tipo test? La opinión de los críticos es que no lo logran.

Considerar las alternativas

El descontento que existe respecto a los exámenes estándar y la profunda fe que se tiene en el valor de la evaluación sistemática, han dado lugar a propuestas que van en pos de nuevas alternativas de evaluación. Ya llamemos a estas alternativas evaluación del rendimiento, evaluación auténtica, evaluación de carpetas de trabajo, evaluación de procesos, exposiciones o demostraciones, lo que se espera de ellas es que consigan resultados educativos significativos y duraderos. Aunque las estrategias de evaluación que se proponen parezcan distintas, todas comparten una visión común (véase la ilustración 1.2).

Ilustración 1.2
Características comunes de las evaluaciones alternativas

- Piden a los alumnos que ejecuten, creen, produzcan o hagan algo.
- Explotan las destrezas más complejas de razonar y de resolver problemas.
- Utilizan tareas que reflejan actividades didácticas significativas.
- Recurren a aplicaciones del mundo real.
- Las personas y no las máquinas son las que evalúan, utilizando así el juicio humano.
- Exigen que los maestros desempeñen un nuevo rol pedagógico y de evaluación.

Además, estos nuevos métodos de evaluación subrayan la importancia de examinar tanto los *procesos* como los productos del aprendizaje. Nos alientan a ir más allá de la idea de que sólo hay "una respuesta correcta" e invitan a los alumnos a explorar las posibilidades intrínsecas de los problemas complejos y sin límite, y a que formen sus propias conclusiones.

La ilustración 1.3 muestra los distintos métodos de evaluación alternativa que se tratan actualmente. Aunque algunas se presentan como nuevas alternativas, en realidad son técnicas de evaluación y aspectos que los maestros vienen tratando desde hace años. Los buenos maestros están siempre pendientes del proceso pedagógico—cómo va una clase, quién tiene problemas, quién presta atención, cómo está trabajando un determinado grupo—y ajustan sus planes didácticos y actividades de acuerdo con éste. De igual manera, la mayoría de los maestros utiliza

una serie de fuentes de información para determinar el progreso de los alumnos en su aprendizaje. Lo nuevo de estas evaluaciones es que vuelven explícito y formal lo que antes era implícito e informal. También alientan a los maestros a articular sus objetivos docentes con claridad, asegurando así una coordinación entre sus objetivos y las teorías actuales de enseñanza efectiva, y recopilar datos sistemáticos para mejorar su rendimiento docente.

Ilustración 1.3
Evaluaciones alternativas

Evaluar procesos	Evaluar productos
• Entrevistas formales	• Ensayos concretos con pautas y criterios de puntuación
• Observaciones documentadas	• Proyectos con criterios de calificación
• Cuadernos y diarios del alumno sobre su aprendizaje	• Carpetas de trabajo de los alumnos con criterios de calificación
• Autoevaluación del alumno (oral o escrita)	• Presentaciones/investigaciones del alumno (expositivas o de expresión artística)
• Entrevistas detalladas sobre proyectos de los alumnos, productos y presentaciones (el alumno explica el qué, el por qué y el cómo, y reflexiona sobre los posibles cambios	• Dibujos, teatro, bailes y cuentos con criterios de calificación
• Listas de control de conducta	• Inventarios de actitud/cuestionarios
• Discusiones entre alumnos junto con exámenes estándar o tipo test	• Exámenes estándar o tipo test, quizás con apartado para las explicaciones

Respaldar la mejora de la calidad docente

La evaluación directa de la expresión escrita del alumno es prueba del poder que podrían llegar a cobrar estos nuevos tipos de evaluación: la integración de la enseñanza y la evaluación. En un distrito los maestros colaboraron para definir los atributos de una buena redacción y para elaborar un baremo de puntuación para medir dichos atributos. Luego capacitaron a otros maestros en la correcta utilización del baremo y se les empleo a éstos en calidad de evaluadores de la expresión escrita en una evaluación a nivel distrito. Los maestros encontraron que los elementos de este baremo constituían una buena base para la enseñanza y ofrecían una forma rápida y uniforme de evaluar y retroalimentar los ejercicios de expresión escrita realizados por los alumnos en el aula. Además, la importancia de la expresión escrita para el distrito, junto con otras iniciativas estatales, alentaron a los maestros a modificar algunos aspectos de la didáctica empleada hasta entonces para esta destreza. El resultado fue una mejora en la expresión escrita por parte de los alumnos y una mayor confianza de los maestros en sus métodos de enseñanza y de evaluación. El desarrollo de exámenes en otras áreas de contenido, que comparten el deseo de renovar los métodos didácticos, es igualmente prometedor.

El camino a seguir en el desarrollo de la evaluación

Si bien la evaluación alternativa implica nuevas formas de interpretar los fines educativos, el procedimiento para desarrollar estas evaluaciones tiene su base en la investigación de sistemas de valoración llevada a cabo durante varias décadas. Aquéllos que van a diseñar exámenes de alta calidad, ya sean exámenes estándar, basados en criterios preestablecidos, o fundamentados en el rendimiento del alumno, se acatan al siguiente procedimiento, si bien con algunas variaciones:

1. Especificar la naturaleza de las destrezas y logros que el alumno tiene que alcanzar.
2. Especificar las tareas ilustrativas que requieren que el alumno demuestre estas destrezas y logros.
3. Especificar los criterios y pautas para valorar el rendimiento del alumno en la tarea.
4. Elaborar un baremo fiable de calificación.
5. Recopilar pruebas de validez para mostrar qué tipos de conclusiones se pueden sacar de la evaluación.
6. Utilizar los resultados de exámenes para refinar la evaluación y para mejorar el currículo y la enseñanza; proporcionar la retroalimentación a los alumnos, los padres y la comunidad.

Los siguientes capítulos describen la manera en la cual se aplica el procedimiento de preparación de exámenes a la evaluación alternativa. El procedimiento se puede modificar según el objetivo de la evaluación, cualquiera que sea su formato. Por ejemplo, en el caso de una evaluación de gran escala o de una evaluación de competencia mínima, donde los resultados son de gran importancia y jugársela a una sola carta es lo habitual, todos estos pasos son imprescindibles. Los pasos cuatro y cinco no serían tan importantes para la evaluación habitual de una clase, donde los maestros disfrutan de múltiples oportunidades para evaluar el progreso de un alumno – sea de manera formal o informal. En el aula, los resultados de cualquier evaluación se ven moderados por otros tipos de datos formales o informales; esto compensa lo que se pierde al no recopilar datos formales de validez y fiabilidad. Sin embargo, los maestros necesitan conocer a fondo las características de algún procedimiento de evaluación técnicamente fiable de manera que puedan desempeñarse en calidad de consumidores hábiles tanto de los productos de evaluación de gran escala como de los productos comerciales que influyen en su didáctica en el aula.

Equilibrar las estrategias de evaluación

No existe una única forma acertada para evaluar a los alumnos. Si bien presentamos un buen argumento para la evaluación del rendimiento, no estamos afirmando que todas las evaluaciones deban ser de este tipo, ni rechazamos el uso de exámenes tipo test u otros exámenes de respuestas preseleccionadas. Lo que afirmamos es que las evaluaciones del rendimiento ofrecen atractivas formas de evaluar las destrezas de razonar y de resolver problemas y, puesto que están fundamentadas en problemas reales, podrían ser más motivadoras para los alumnos y reforzarían más sus conocimientos. Sin embargo, mientras que las evaluaciones del rendimiento pueden proporcionar información sobre el éxito de los alumnos en cuanto a la aplicación de sus conocimientos, los exámenes tipo test pueden demostrar ser más eficaces para determinar el grado de adquisición de conceptos e información básica por parte de los alumnos. Un currículo equilibrado requiere de un método equilibrado de evaluación.

Además, no sólo porque una evaluación requiera que el alumno lleve a cabo una actividad interesante o compleja significa que sea una buena evaluación. Una buena evaluación *mide* objetivamente algo más que las tareas específicas que se pide de los alumnos. Los resultados de una buena evaluación identifican lo que pueden hacer los alumnos en un amplio dominio de conocimientos o destrezas. Las destrezas que exhiben los alumnos en el contexto de la evaluación deben poderse trasladar a otras situaciones y problemas.

Mantener un nivel alto en las evaluaciones

Sin tener en cuenta su finalidad o formato, las evaluaciones de calidad deben reunir ciertas condiciones comunes. El Centro para la Investigación de la Evaluación, Criterios y Examinación del Alumno (The Center for Research on Evaluation, Standards and Student Testing; CRESST), (Linn, Baker y Dunbar, 1991) ha establecido criterios clave para un procedimiento completo de elaboración de una evaluación. Entre estos criterios se incluyen los siguientes:

- **Consecuencias.** Los antecedentes del proceso de examen están colmados de ejemplos de buenas intenciones que han fracasado. Este criterio requiere de una planificación desde el principio para evaluar las verdaderas consecuencias de la evaluación. ¿Tiene consecuencias positivas o hay efectos no intencionados como es la limitación del currículo, efectos negativos para los alumnos más atrasados, etcétera?
- **Objetividad.** ¿Toma en cuenta la evaluación el historial cultural de esos alumnos que se examinan? ¿Han tenido todos los alumnos las mismas oportunidades de aprender las destrezas de razonar y de resolver problemas que se están evaluando?
- **Transferencia y Generalización.** ¿habrá concordancia entre los resultados de la evaluación y las generalizaciones que se han establecido acerca de la capacidad de los alumnos? ¿Son fiables los resultados dados por distintos evaluadores y tienen el mismo significado en distintas localidades?
- **Complejidad Cognitiva.** No se puede averiguar a simple vista si una evaluación evalúa realmente o no las destrezas de razonar. ¿*Requiere* de hecho una evaluación que los alumnos utilicen la destreza de razonar y solucionar problemas?
- **Calidad de contenido.** Los ejercicios seleccionados para tener una idea del dominio de un contenido en particular deben concordar con el tiempo y el esfuerzo que los alumnos y evaluadores han invertido. ¿Es el contenido seleccionado consistente con la mejor y más actual investigación de este campo y refleja los aspectos importantes de una disciplina que seguirá siendo válida?
- **Contenido.** El criterio de selección del contenido de una disciplina requiere que la evaluación concuerde con el currículo y, a lo largo de una serie de evaluaciones, que represente al currículo en su totalidad. Debido a las restricciones de tiempo que van a limitar el número de evaluaciones alternativas que se podrían realizar, un contenido adecuado representa un reto importante. ¿Se ha incluido los elementos clave del currículo en esta serie de evaluaciones?
- **Valor Significativo.** Una de las razones para justificar la utilización de evaluaciones de mayor contexto es que aseguran que los alumnos se vean obligados a enfrentarse a problemas que tienen sentido y que resultan ser experiencias educativas útiles y más motivadoras. ¿Consideran los alumnos que los ejercicios de evaluación son realistas y útiles?

- **Costo y Eficacia.** Para que sean eficaces, las evaluaciones deben ser económicas. Las evaluaciones de trabajo intensivo basadas en el rendimiento requieren de una recopilación de datos y sistemas de puntuación eficaces. ¿Merece esta información sobre los alumnos el costo y el tiempo que lleva obtenerla?

Finalmente, es importante apuntar que la evaluación alternativa es un campo en desarrollo. Nuevas estrategias, así como nuevas metodologías, están evolucionando para asegurar su buena calidad. Mientras más aprendemos sobre la evaluación alternativa, se pueden refinar o incluso volver a formular los métodos actuales.

Referencias bibliográficas

Baker, E.L. (1989). "Mandated Tests: Educational Reform or Quality Indicator?" En *Test Policy and Test Performance: Education, Language, and Culture*, redactor B.R. Gifford. (pp. 3-23). Norwell, Mass.: Kluwer.

Cannel, J.J. (1987). *Nationally Normed Elementary Achievement Testing in America's Public Schools: How all 50 States Are Above the National Average*. (2ª ed.). Daniels, W. Va.: Friends of Education.

Herman J. y S. Golan. (1990). *Effects of Standardized Testing on Teachers and Learning: Another Look*. (Tech. Rep. No. 334). Los Angeles: University of California, Center for the Study of Evaluation.

Linn, R.L., E.L. Baker, y S.B. Dunbar. (1991). "Complex, Performance-based Assessment: Expectations and Validation Criteria." *Educational Researcher* 20, 8: 15-23.

Linn, R.L., M.E. Graue, y N.M. Sanders. (1990). *Comparing State and District Test Results to National Norms: Interpretations of Scoring "Above the National Average."* (CSE Tech. Rep. No. 308). Los Angeles: University of California, Center for the Study of Evaluation.

Shepard, L.A. (Abril 1989). "Why We Need Better Assessments." *Educational Leadership* 46, 7: 4-9.

2

Vincular la evaluación y la enseñanza

Las nuevas perspectivas de un currículo, de una enseñanza y de un aprendizaje eficaces requieren de un nuevo enfoque de la evaluación sistemática. Ya no se concibe el aprendizaje como una transmisión unilateral del maestro hacia los alumnos con la ilustración del maestro como transmisor y los alumnos como receptores pasivos. Por el contrario, la enseñanza significativa implica la participación activa de los alumnos en el proceso de aprendizaje. Los buenos maestros extraen y sintetizan el conocimiento de la disciplina, del aprendizaje del alumno, y del desarrollo infantil. Utilizan una amplia gama de estrategias pedagógicas, que abarcan desde la enseñanza directa a la particularizada, con el objeto de lograr la participación de los alumnos en actividades significativas—debates, tareas en equipo, proyectos prácticos—y de lograr objetivos específicos de aprendizaje. Los buenos maestros evalúan constantemente el avance de sus alumnos, recopilan información sobre sus problemas y su progreso, y de acuerdo a ello van modificando su plan didáctico.

En este capítulo analizaremos las tendencias educativas y sociales que respaldan estas nuevas perspectivas de la enseñanza y del aprendizaje que han suscitado la necesidad de nuevas técnicas de evaluación (véase ilustración 2.1). Estas mismas tendencias exigen más que nunca una gran preparación por parte de los maestros , ya que deben integrar el conocimiento de los objetivos proyectados, los procesos de aprendizaje, los contenidos del currículo, y la evaluación.

Ilustración 2.1
Nuevas tendencias de evaluación

1. Cambio del enfoque conductista sobre el aprendizaje al enfoque cognitivo sobre la evaluación
   - Del énfasis exclusivo en el producto o el resultado del aprendizaje del alumno a la preocupación por el proceso de aprendizaje
   - De la actitud pasiva a la construcción activa del significado
   - De la evaluación de destrezas por separado y aisladas a la evaluación integrada y multidisciplinaria
   - Énfasis en la metacognición (autorregulación y destrezas del aprendizaje a la habilidad de aprender) y destrezas conativas (la motivación y otras áreas afectivas que influyen en el aprendizaje y el éxito escolar)
   - Cambio de lo que se entiende por saber y estar bien preparado—de la acumulación de hechos y destrezas aisladas a la importancia de la aplicación y del uso del conocimiento

2. De la evaluación basada en exámenes escritos a la evaluación auténtica
   - Pertinente y significativo para los alumnos
   - Problemas contextualizados
   - Énfasis en las destrezas complejas
   - No hay una única respuesta correcta
   - Criterios públicos conocidos anteriormente
   - Avance individual y desarrollo

3. Carpetas de trabajos: de la evaluación en una sola instancia a muestras de trabajo a largo plazo
   - Fundamento de la evaluación por el profesor
   - Fundamento de la autoevaluación por los alumnos
   - Fundamento de la evaluación por los padres

4. De la evaluación de un sólo atributo a una evaluación multidimensional
   - Reconocimiento de las múltiples destrezas y aptitudes de los alumnos
   - Mayor sensibilidad a la maleabilidad de la capacidad del alumno
   - Oportunidades para que el alumno desarrolle y demuestre diversas habilidades

5. De un énfasis casi exclusivo en la evaluación individual a la evaluación colectiva
   - Destrezas de trabajar en equipo
   - Productos en colaboración

Enfrentarse a las nuevas exigencias de la educación

Consideremos lo que las predicciones futuristas implican para los objetivos educativos y para los tipos de destrezas que los alumnos y la sociedad entera van a necesitar en el siglo XXI (Benjamin 1989). El conocimiento está en expansión geométrica; la base del conocimiento mundial se ha cuadriplicado en este siglo (Cornish 1986). Dada esta velocidad, no se puede esperar de ningún individuo que se mantenga al día con toda la nueva información que surge en una sola disciplina, ni mucho menos en varias. Debido a esta explosión de conocimiento, casi todos los esfuerzos por conseguir que los alumnos memoricen y reproduzcan grandes cantidades de información son inútiles.

Las tendencias económicas actuales también nos alejan de un currículo basado en hechos. El cambio de una economía de manufactura a una economía basada en la informática y los servicios requiere que los individuos desarrollen destrezas de utilizar y acceder a información y que adquieran destrezas para trabajar en equipo. Estos cambios en el mundo profesional y el ritmo y complejidad de la vida moderna sugieren que vamos a tener que ser flexibles, cambiar de trabajo frecuentemente y adaptarnos a los cambios. En la formación de los alumnos para su éxito profesional, las escuelas deben enseñar a manejar información en lugar de simplemente enseñar a adquirirla.

Utilizar las teorías cognitivas de aprendizaje

Las nuevas teorías cognitivas del aprendizaje nos llevan por caminos similares. Las primeras teorías de aprendizaje suponían que las destrezas complejas se adquirían poco a poco en una secuencia cuidadosamente estructurada de destrezas más sencillas. Éstas eran requisitos previos y parte de dicha secuencia, y se articulaban con frecuencia en objetivos de conducta aislados. Se suponía que se debía enseñar las destrezas básicas necesarias para el aprendizaje mecánico antes de continuar con destrezas superiores que requerían de un aprendizaje por descubrimiento. Sin embargo, datos procedentes de la psicología cognitiva indican que el aprendizaje no es lineal y que tampoco se consigue uniendo fragmentos de un aprendizaje más sencillo. El aprendizaje es un proceso continuo durante el cual los alumnos están constantemente recibiendo información, interpretándola, incorporándola a sus conocimientos y experiencias (sus conocimientos previos) y reorganizando y revisando el concepto que se han formado del mundo, llamadas "modelos mentales", "estructuras de conocimientos" o "esquema".

La naturaleza activa del aprendizaje

Visto desde la perspectiva contemporánea cognitiva, el aprendizaje significativo es reflexivo, constructivo y de autorregulación (Wittrock 1991, Bransford y Vye 1989, Marzano et al. 1988, Davis et al. 1990). No simplemente grabamos información de

los hechos, sino que creamos nuestra propia interpretación del mundo—nuestras propias estructuras de conocimientos. *Saber* algo no es sólo recibir información de forma pasiva, sino interpretarla e incorporarla a nuestros conocimientos previos. Además, hoy reconocemos la importancia de saber no sólo cómo actuar sino también cuándo actuar y cómo adaptar esa actuación a situaciones nuevas. La presencia o ausencia de fragmentos aislados de información, que suele ser el eje central de muchos de los exámenes tipo test, no es de suma importancia en la evaluación de un aprendizaje significativo. En su lugar, lo que nos interesa es saber si los alumnos organizan, estructuran y utilizan esa información en un contexto en el que tienen que resolver problemas complejos y averiguar cómo lo hacen.

El aprendizaje no es lineal

El aprendizaje no consiste en jerarquías aisladas. Puesto que el aprendizaje no es lineal y puede tomar varias direcciones simultáneamente a un ritmo irregular, el aprendizaje de conceptos no es algo que pueda retrasarse a una edad en particular o hasta que se haya dominado todos los "hechos básicos". Personas de todas las edades y con diferentes capacidades utilizan y refinan conceptos constantemente.

Hoy en día tenemos pruebas que dejan claro que la enseñanza que se concentra en prácticas repetitivas estructuradas (*drills*) y en la práctica de hechos y destrezas aisladas perjudica a los alumnos. Insistir que los alumnos demuestren cierto nivel en el dominio de las matemáticas antes de dejarles estudiar álgebra o que aprendan a redactar un párrafo debidamente antes de intentar redactar un ensayo son ejemplos de este método basado en destrezas aisladas. Este tipo de aprendizaje fuera de contexto hace que sea más difícil organizar y recordar la información que se presenta. De igual manera resulta difícil aplicar las destrezas que se han enseñado en el aula a la resolución de problemas en el mundo real. Los alumnos que tienen problemas con el dominio de estos "conceptos básicos" fuera de contexto, se ven frecuentemente relegados a clases o grupos de recuperación sin que se les brinde la oportunidad de abordar tareas complejas y significativas.

Los alumnos tienen múltiples aptitudes

Las teorías actuales sobre la inteligencia destacan la existencia de una amplia gama de talentos y capacidades humanas, y no están de acuerdo con la opinión popular de que la inteligencia o habilidad consiste en una capacidad única y fija (Sternberg 1991, Gardner 1982). Gardner señala que mientras la educación tradicional ha dado importancia a tan sólo dos habilidades, la verbal-lingüística y la lógica-matemática, existen también muchas más "inteligencias" importantes, como son la visual-espacial, la quinestésica, la musical, la intrapersonal y la interpersonal. Gardner afirma que todos los individuos tenemos fuerzas en dos o tres de estas áreas. Además, los modos y velocidades que empleamos cuando adquirimos conocimientos son muy diversos, al igual que las capacidades de atención y

memoria que podemos utilizar en la adquisición de conocimientos y en la actuación, y las diferentes formas de demostrar los distintos significados que hemos creado. Para lograr el éxito con todos los alumnos, la enseñanza y la evaluación necesitan aprovechar algo más que las inteligencias lingüísticas o la lógica-matemática y admitir la premisa de que todos los alumnos son capaces de aprender.

El aprendizaje incluye la cognición, la metacognición y el afecto

Los estudios recientes sobre la integración del aprendizaje y de la motivación destacan la importancia de las destrezas afectivas y metacognitivas (pensar sobre pensar) en el aprendizaje (McCombs 1991, Weinstein y Meyer 1991). Por ejemplo, Belmont et al. (1982) sugieren que la diferencia entre los que razonan y resuelven problemas con dificultad y los que lo hacen bien no se encuentra simplemente en las destrezas que poseen, sino en su no utilización. La simple adquisición de conocimientos y destrezas no vuelve más capaces a los individuos en lo referente a pensar y resolver problemas. También deben adquirir la costumbre de emplear las destrezas y estrategias y saber cuándo aplicarlas.

La investigación y la experiencia, como las que se han llevado a cabo en el campo de la expresión escrita (Gere y Stevens 1985, Burnham 1986), demuestran la importancia de hacer reflexionar a los alumnos sobre lo que constituye un trabajo excelente y sobre cómo evaluar sus propios esfuerzos. Si facilitamos a los alumnos modelos de actuación ejemplar y les animamos a reflexionar sobre sus trabajos, les ayudamos a entender e interiorizar nuestros criterios.

El aprendizaje significativo se considera una motivación intrínseca. El valor a largo plazo de los tradicionales motivadores extrínsecos, como son las notas o los premios, es discutible. La investigación indica que estas técnicas pueden incluso desvalorizar la motivación intrínseca del alumno, lo que influye negativamente en el dominio o rendimiento escolar (Lepper y Greene 1978).

El contexto social del aprendizaje

En los últimos años también se ha prestado atención al papel que juega el contexto social al dar forma a las habilidades y disposiciones cognitivas complejas. Aunque los problemas de la vida real muchas veces nos obligan a trabajar en equipo, la mayor parte de la enseñanza y evaluación tradicional se ha basado en el trabajo individual. Hoy en día sabemos que el trabajo en equipo facilita el aprendizaje. Trabajar junto con compañeros en una tarea en común proporciona: (1) muchos modelos de estrategias de razonamiento eficaces; (2) la retroalimentación mutua constructiva; (3) el reconocimiento de la importancia de colaborar con otros; y (4) ayuda para alcanzar destrezas o conocimientos difíciles o complejos.

Las exigencias de una democracia proporcionan otras razones fundamentales que respaldan la importancia de la investigación en equipo. Se espera que los alumnos que trabajan juntos en una "comunidad de estudiantes" se escuchen unos

a otros con respeto, reflexionen y crezcan con las ideas de los demás, exijan pruebas para apoyar las opiniones de los otros, se ayuden a la hora de sacar conclusiones y cuestionen los hechos, suposiciones y argumentos de diferentes puntos de vista (Jones y Fennimore 1990).

Centrarse en un currículo razonado

Un método moderno sobre cómo hacer un currículo, llamado "Currículo Razonado" (*Thinking Curriculum*) diseñado por Lauren Resnick y Leopold Klopfer (1989), recomienda firmemente una visión integrada y activa del aprendizaje del alumno. El currículo razonado destaca la importancia tanto del proceso como del producto. Muchas veces los alumnos realizan tareas parecidas a aquéllas que encuentran en el mundo real. Los alumnos llevan a cabo tareas que requieren de razonamiento complejo, planificación y evaluación. Resuelven problemas, toman decisiones, construyen argumentos, etcétera. De este modo, imitan el proceso de una disciplina profesional a la vez que adquieren conocimientos de esa disciplina.

Según Fennimore y Tinzmann (1990), los cuatro principios clave que se detallan a continuación caracterizan un currículo razonado.

La promoción de un aprendizaje profundo

Un currículo razonado ayuda a los alumnos a adquirir los conceptos y herramientas clave para crear, utilizar y comunicar los conocimientos en un determinado campo. El suficiente conocimiento de un campo implica una red integrada de conocimientos y conceptos en lugar de una recopilación de hechos aislados.

En un currículo razonado los alumnos desarrollan una capacidad de comprensión profunda de los conceptos esenciales y de los procesos que necesitan para enfrentarse a estos conceptos, parecida a los métodos utilizados por los expertos al abordar sus tareas. Por ejemplo, los alumnos utilizan fuentes primarias para construir hechos históricos; diseñan experimentos para dar respuesta a sus preguntas sobre fenómenos naturales; utilizan las matemáticas en relación con los sucesos y sistemas del mundo real; y escriben para un lector real.

Objetivos de contenido y proceso en tareas del mundo real

En lugar de centrarse en destrezas simples y aisladas, los alumnos adoptan el razonamiento complejo e integral para enfrentarse a cuestiones fuera del aula. Según Resnick (1989) este razonamiento de la vida real muchas veces implica: procesos significativos de tomar decisiones y resolver problemas; colaborar con otros; utilizar las herramientas disponibles; establecer lazos con los sucesos y objetos del mundo real; y utilizar conocimientos interdisciplinarios.

Desempeño integral en entornos cada vez más desafiantes

Un currículo razonado no aísla destrezas y hechos. En su lugar incluye el desempeño integral de tareas significativas y complejas en entornos cada vez más desafiantes. Los materiales y el contenido están estructurados de tal manera que los alumnos regulan su propio aprendizaje de forma gradual. Este método asegura que el aprendizaje motive a los alumnos y despierta en ellos un sentido de eficacia y confianza.

Concatenar el contenido y el proceso a la experiencia previa de los alumnos

Un currículo razonado tiene en cuenta las experiencias y los conocimientos que el alumno lleva consigo cuando va a la escuela. Luego se aumenta y refina estos conocimientos previos al concatenarlos con el nuevo aprendizaje. Esto logra que el contenido curricular sea pertinente a los asuntos y tareas importantes en la vida de los alumnos. Cuando los alumnos relacionan el aprendizaje escolar con su vida real están más dispuestos a buscar y valorar las perspectivas de los otros—compañeros, maestros, padres, miembros de su comunidad y expertos. Al hacerlo así, desarrollan competencias interpersonales para crear y participar en diálogos con individuos que poseen diferentes perspectivas y que proceden de distintos ambientes.

Vincular la evaluación y la enseñanza

En la ilustración 2.2 se resume muchos de los principios básicos del aprendizaje que hemos tratado en este capítulo y se describe algunas de las consecuencias que estos principios tienen tanto en la enseñanza como en la evaluación. Como muestra la ilustración 2.2, la evaluación no sólo evalúa cuánto se ha aprendido en una determinada unidad didáctica, sino que además proporciona información actualizada a alumnos y maestros sobre su progreso y posibles formas de mejorar.

Ilustración 2.2
Vincular la evaluación y la enseñanza: Implicaciones de la teoría cognitiva del aprendizaje

Teoría: El conocimiento se construye. El aprendizaje es un proceso de crear un significado personal utilizando la información nueva y los conocimientos previos.

Implicaciones para la enseñanza/evaluación:

- Fomentar la discusión de ideas nuevas.
- Fomentar el razonamiento divergente, eslabones y soluciones múltiples, no sólo una respuesta correcta.
- Fomentar modos de expresión múltiples, por ejemplo, caracterizaciones, simulacros, debates y exposiciones.
- Entatizar las destrezas de pensamiento crítico: analizar, comparar, generalizar, predecir, formar hipótesis.
- Relacionar la nueva información con la experiencia personal, conocimientos previos.
- Aplicar información a una situación nueva.

Teoría: Todas las edades/capacidades pueden razonar y resolver problemas. El aprendizaje no sigue necesariamente una progresión lineal de destrezas aisladas.

Implicaciones para la enseñanza/evaluación:

- Lograr la participación de todos los alumnos en la resolución de problemas.
- No hacer que la resolución de problemas, el pensamiento crítico o la discusión de los conceptos dependa del dominio de las destrezas básicas habituales.

Teoría: Hay una gran variedad de estilos de aprendizaje, períodos de concentración, memorias, velocidades del desarrollo e inteligencias.

Implicaciones para la enseñanza/evaluación:

- Ofrecer opciones de tareas (no sólo la comprensión de la lectura y la expresión escrita).
- Ofrecer opciones de cómo demostrar dominio/aptitud.
- Dar tiempo para planificar y realizar trabajos.
- No explotar el uso de exámenes cronometrados.
- Ofrecer la oportunidad de revisar y volver a pensar.
- Incluir experiencias concretas (manipulativas, vínculos con experiencias previas).

(continuación)

Ilustración 2.2—*continuación*

Teoría: Los individuos rinden mejor cuando conocen el objetivo, ven modelos, interpretan su rendimiento comparándolo con la norma.

Implicaciones para la enseñanza/evaluación:

- Discutir objetivos; permitir que los alumnos ayuden a definirlos (los personales y los del aula).
- Proporcionar una amplia gama de muestras de trabajos de alumnos; discutir las características.
- Proporcionar a los alumnos la oportunidad de autoevaluarse y de evaluarse entre compañeros.
- Discutir los criterios para calificar el rendimiento.
- Permitir que los alumnos ofrezcan ideas sobre los criterios.

Teoría: Es importante saber cuándo utilizar los conocimientos, cómo adaptarlos, cómo dirigir el propio aprendizaje.

Implicaciones para enseñanza/evaluación:

- Ofrecer oportunidades reales (o simulacros) para aplicar/adaptar nuevos conocimientos.
- Hacer que los alumnos se autoevalúen: pensar sobre cómo aprenden bien/mal; establecer nuevas metas, por qué les gustan ciertos ejercicios.

Teoría: La motivación, el esfuerzo y la autoestima influyen en el aprendizaje y en el rendimiento.

Implicaciones para la enseñanza/evaluación:

- Motivar a los alumnos con tareas reales y vínculo con experiencias personales.
- Ayudar a los alumnos a que aprecien el vínculo entre esfuerzos y resultados.

Teoría: El aprendizaje tiene componentes sociales. El trabajo en equipo tiene mucho valor.

Implicaciones para la enseñanza/evaluación:

- Incluir el trabajo en equipo.
- Incorporar grupos heterogéneos.
- Dejar que los alumnos desempeñen papeles distintos.
- Considerar los productos de grupo y los procesos de grupo.

Los diferentes tipos de evaluación fomentan objetivos múltiples que incluyen, pero no de forma exclusiva, la adquisición de conocimientos de contenido curricular. Los exámenes ya no se limitan a tareas escritas, cronometradas y preestablecidas para que los alumnos en solitario demuestren lo que saben. La evaluación hoy en día se realiza en muchos contextos e incluye el trabajo individual y en equipo, ejercicios con ayuda o sin ella, y períodos de tiempo cortos

o largos. La libre discusión entre maestros, alumnos e incluso padres sobre los criterios de rendimiento y calificaciones que se han de utilizar, es la marca distintiva de la evaluación alternativa. Al ser la evaluación una parte integral de la enseñanza, la consideración de objetivos docentes es el primer paso crucial en la planificación de tareas de evaluación y baremos de puntuación significativos.

Referencias bibliográficas

Belmont, J.M., E.C. Butterfield, y R.P. Ferretti. (1982). "To Secure Transfer of Training, Instruct Self-management Skills." En *How and How Much Can Intelligence Be Increased?* redactores D.K. Detterman y R.J. Sternberg. Norwood, N.J.: Ablex.

Benjamin, S. (1989). "An Ideascape for Education: What Futurists Recommend." *Educational Leadership* 7, 1:8-14.

Bransford, J.D., y N. Vye. (1989). "A Perspective on Cognitive Research and its Implications in Instruction." En *Toward the Thinking Curriculum: Current Cognitive Research (1989 Yearbook of the Association for Supervision and Curriculum Development)*, redactores L.B. Resnick y L.E. Klopfer. Alexandria, Va.: Association for Supervision and Curriculum Development.

Burnham, C. (1986). "Portfolio Evaluation: Room to Breathe and Grow." En *Training the Teacher of College Composition*, redactor C. Bridges. Urbana, Ill.: National Council of Teachers of English.

Cornish, E. (1986). "Educating Children for the 21st Century." *Curriculum Review* 25, 4: 12-17.

Davis, R.B., y C.A. Maher. (1990). "Constructivist View of the Teaching of Mathematics." *Journal for Research in Mathematics Education*. Reston, Va.: National Council of Teachers of Mathematics.

Fennimore, T.F., y M.B. Tinzmann. (1990). "Restructuring to Promote Learning in America's Schools: Video Conference 2: The Thinking Curriculum." Elmhurst, Ill.: North Central Regional Educational Laboratory.

Gardner, H. (1982). *Art, Mind and Brain*. New York: Basic Books.

Gere, A., y R. Stevens. (1985). "The Language of Writing Groups: How Oral Response Shapes Revision." En *The Acquisition of Written Language: Response and Revision*, redactor S.W. Freedman.

Jones, B.F., y T.F. Fennimore. (1990). *The New Definition of Learning: The First Step to School Reform*. Chicago: North Central Regional Educational Laboratory.

Lepper, M.R., y D. Greene. (1978). *The Hidden Costs of Reward: New Perspectives on the Psychology of Human Motivation*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Marzano, R., R. Brandt, y C.S. Hughes. (1988). *Dimensions of Thinking: A Framework for Curriculum and Instruction*. Alexandria, Va.: Association for Supervision and Curriculum Development.

McCombs, B.L. (1991). "The Definition and Measurement of Primary Motivational Processes." En *Testing and Cognition*, redactor M.C. Wittrock y E.L. Baker. Englewood Cliffs, N.J.: Prentince-Hall.

Resnick, L.B., y L.E. Klopfer. (1989). "Toward the Thinking Curriculum: An Overview." En *Toward the Thinking Curriculum: Current Cognitive Research (1989 Yearbook of the Association for Supervision and Curriculum Development)*, redactores L.B. Resnick y L.E. Klopfer. Alexandria, Va.: ASCD.

Sternberg, R.J. (1991). "Toward Better Intelligence Tests." En *Testing and Cognition*, redactores M.C. Wittrock y E.L. Baker. Englewood Cliffs, N.J.: Prentice-Hall.

Weinstein, C.E., y D.K. Meyer. (1991). "Implications of Cognitive Psychology for Testing: Contributions from Work in Learning Strategies." En *Testing and Cognition*, redactores M.C. Wittrock y E.L. Baker. Englewood Cliffs, N.J.: Prentice-Hall.

Wittrock, M.C. (1991) "Testing and Recent Research in Cognition." En *Testing and Cognition*, redactores M.C. Wittrock y E.L. Baker. Englewood Cliffs, N.J.: Prentice-Hall.

3

Establecer los objetivos

El primer paso en la elaboración o selección de una evaluación es conocer los objetivos de esa evaluación: ¿para qué se necesitan los resultados? ¿Qué aspectos del rendimiento del alumno interesan?

Aunque este libro no pretende ser un manual sobre los objetivos y aplicaciones de la evaluación, habrá que tener en cuenta cuáles son los objetivos que se desean a lo largo de todo el proceso evaluativo. ¿Es nuestro objetivo principal evaluar los logros de los alumnos—por ejemplo, qué tal han aprendido a escribir relatos, a comunicarse oralmente, a sintetizar su investigación? Si es así, nuestro mayor interés será evaluar la calidad o nivel de los logros de los alumnos a efectos de calificación, de colocación por niveles y de seguimiento del progreso, o a efectos de objetivos de evaluación y responsabilidad adjudicada o rendir cuentas (*accountability*) a nivel escuela, distrito y a nivel de otros objetivos extraescolares. Puesto que la principal finalidad es averiguar hasta qué punto los alumnos han logrado asimilar distintos conocimientos y destrezas, la evaluación debe enfocarse en los resultados o el producto del aprendizaje del alumno.

Sin embargo, si el objetivo de la evaluación es diagnosticar para luego mejorar, por ejemplo el diagnóstico de los puntos fuertes y débiles de un alumno, recomendar los programas de enseñanza más apropiados, o identificar las estrategias que los alumnos saben emplear bien y aquellas con las que tienen problemas, necesitaremos una evaluación que nos proporcione información tanto

del proceso como del resultado. ¿Qué han logrado los alumnos y cómo lo han hecho? La información de procesos nos proporciona tales explicaciones.

Los objetivos y la utilización de una evaluación determinada están directamente relacionados con el tiempo que dedicamos a la recopilación de pruebas de fiabilidad y validez, tema que tratamos más detalladamente en los capítulos 6 y 7. Entre más se ponga en juego todo lo referente a la evaluación, mayor será la necesidad de recopilar informes de fiabilidad y validez. Es fundamental adecuar los niveles de ambos cuando, por ejemplo, los resultados van a utilizarse para pasar a un alumno a un curso superior, o transferirlo a una clase de atención especial, o para premiar a maestros o escuelas.

Establecer los objetivos didácticos principales

Una buena evaluación exige que conozcamos y seamos capaces de articular nuestros objetivos didácticos principales. Estos determinan qué aspectos del rendimiento nos interesa conocer. ¿Qué es lo que queremos que los alumnos logren en una unidad didáctica, un curso, una disciplina, o en varias disciplinas? ¿En qué debería consistir nuestro programa didáctico? ¿Qué deberían ser capaces de hacer los alumnos al término de una unidad didáctica, de un curso o de un año escolar que no podían hacer antes? ¿Cuáles son las áreas importantes del desarrollo del alumno sobre las que queremos influir?

Las respuestas a estas preguntas definen nuestras prioridades en el aula y representan los principales objetivos de nuestras actividades didácticas. Estas mismas prioridades deberían también ser la base de las tareas de evaluación que requerimos de los alumnos. Esto contribuye a una evaluación justa—los alumnos tienen la oportunidad de adquirir los conocimientos y destrezas que estamos evaluando—y también contribuye a una tarea de evaluación significativa que refuerza las destrezas y logros que consideramos más importantes.

Establecer nuestras finalidades prioritarias

Mientras que designar objetivos puede parecer fácil, es desafiante establecer prioridades de entre miles de posibilidades. ¿Qué campos del conocimiento, destrezas y disposiciones valen la pena enseñar y evaluar? ¿Qué finalidades educativas se pretende alcanzar? Puesto que las evaluaciones del rendimiento requieren de mucho tiempo y energía—tanto por nuestra parte como por parte de nuestros alumnos—habrá que centrarse en un número relativamente pequeño de finalidades importantes, pudiendo cada una representar un mes o un trimestre de docencia. Estas evaluaciones deben enfocarse en nuestros principales objetivos de aprendizaje para nuestros alumnos. Para poder definir estos objetivos, es indispensable formularse las siguientes preguntas interrelacionadas (a las que se ha añadido algunas posibles respuestas):

1. ¿Qué destrezas cognitivas importantes deseo que adquieran los alumnos?
Quiero que los alumnos sean capaces de:

- Comunicarse por medio de la expresión escrita de forma eficaz, o para ser más preciso, escribir de una manera persuasiva, escribir buenas descripciones y escribir relatos.
- Comunicarse oralmente con eficacia.
- Analizar la literatura teniendo en cuenta el argumento, los personajes, la ambientación, y el tema.
- Analizar temas utilizando fuentes primarias y material de consulta.
- Utilizar el álgebra para solucionar problemas cotidianos.
- Analizar acontecimientos actuales desde perspectivas históricas, políticas, geográficas y multiculturales.
- Diseñar y dirigir estudios para facilitar la toma de decisiones sobre problemas actuales o cotidianos.
- Utilizar el método científico.
- Utilizar distintos medios para expresar lo que saben.

2. ¿Qué destrezas sociales y afectivas deseo que adquieran los alumnos?
Quiero que sean capaces de:

- Trabajar de forma independiente.
- Desarrollar un espíritu de equipo y las destrezas de trabajo en equipo.
- Apreciar sus propios puntos fuertes.
- No desistir ante los desafíos.
- Enorgullecerse de su trabajo.
- Disfrutar y valorar el aprendizaje.
- Confiar en sus habilidades.
- Tener un escepticismo sano acerca de las polémicas y posturas de la actualidad.
- Comprender que todos tenemos nuestros puntos fuertes y que cualquier persona puede destacar de alguna forma.

3. ¿Qué destrezas metacognitivas deseo que adquieran los alumnos?
Quiero que sean capaces de:

- Reflexionar sobre el proceso de escribir que emplean, evaluar su eficacia y formular sus propias ideas para mejorarlo.
- Discutir y evaluar sus estrategias para la resolución de problemas.
- Formular planes eficaces para completar sus proyectos individuales y para controlar su progreso.
- Evaluar la eficacia de sus estrategias de investigación.

4. ¿Qué tipos de problemas deseo que los alumnos sean capaces de resolver?
Quiero que:

- Sepan cómo investigar.
- Resuelvan problemas que requieren de demostración geométrica.
- Comprendan los tipos de problemas que la trigonometría les puede ayudar a resolver.
- Apliquen el método científico.
- Predigan consecuencias.

- Resuelvan problemas que no tienen una única respuesta correcta.
- Tomen decisiones apropiadas.
- Creen expresiones propias originales.

5. ¿Qué conceptos y principios deseo que los alumnos sean capaces de aplicar?
Quiero que sean capaces de:

- Entender el concepto de democracia.
- Entender las relaciones causa-efecto en la historia y en la vida cotidiana.
- Entender el significado de diversas propuestas lógicas.
- Criticar obras literarias basándose en su argumento, ambientación, intención, etcétera.
- Entender y reconocer las consecuencias del abuso de estupefacientes.
- Aplicar los principios básicos de la ecología y de la conservación en su vida diaria.

Debemos ser lo más específicos posible a la hora de formular las respuestas a estas preguntas. No debiéramos aspirar a una información tan detallada como la que se utilizó en los objetivos de conducta del pasado, sino que debemos describir nuestras finalidades prioritarias con suficiente detalle como para que otros puedan estar de acuerdo con lo que significan estas finalidades y si es que los alumnos las han alcanzado.

Utilizar fuentes disponibles

Además de utilizar nuestra propia opinión a la hora de contestar las preguntas que acabamos de formular, también podría resultar útil consultar directrices curriculares, expertos en contenido curricular o proyectos innovadores que reflejen la filosofía educativa individual. A continuación se detallan algunas fuentes que podrían ser de interés.

Grupos curriculares nacionales

Una fuente de gran utilidad es el *Curriculum and Evaluation Standards for School Mathematics*, publicado por el "Consejo Nacional de Maestros de Ciencias Exactas" ("National Council of Teachers of Mathematics"; 1989). En esta fuente se da gran importancia al desarrollo de las capacidades de los alumnos para la utilización de las matemáticas a la hora de resolver problemas, razonar y comunicarse. Además, alienta a los alumnos a valorar las matemáticas y a sentirse seguros sobre sus habilidades matemáticas. Por ejemplo, según los criterios de comunicación del "NCTM" los alumnos deberían ser capaces de:

- Articular las razones por las que utilizan una representación o solución matemática en particular:

- Interpretar y resumir los datos que han recopilado;
- Describir como se relacionan los conceptos matemáticos con los modelos físicos o gráficos; y
- Justificar sus argumentos utilizando el razonamiento deductivo o inductivo.

Estos principales objetivos del rendimiento del alumno pueden estimular las ideas sobre aquellos objetivos que se quiere establecer para los alumnos en el campo de las matemáticas.

De hecho, grupos de distintas disciplinas académicas están desarrollando, o han desarrollado sus propias listas de objetivos. La "Asociación Americana para el Avance de la Ciencia" ("American Association for the Advancement of Science;" AAAS 1989) ha formulado una serie de recomendaciones para la reestructuración del currículo de las ciencias que aparecen en la publicación titulada *Science for All Americans: Project 2061*. Este informe recomienda cuatro objetivos para la enseñanza de las ciencias: entender el esfuerzo científico, desarrollar visiones científicas del mundo, formular perspectivas históricas y sociales de la ciencia y adquirir hábitos mentales científicos.

El "Consejo Nacional de Maestros de Ciencias Sociales" ("National Council of Teachers of Social Studies"), el "Consejo Nacional de Maestros de Ciencias" ("National Council of Teachers of Science") y el "Consejo Nacional de Maestros de Inglés" ("National Council of Teachers of English") son todos fuentes valiosas de información en sus disciplinas respectivas. El "Centro de Educación Cívica" ("Center for Civic Education") ha publicado *Civitas* que cubre toda la educación cívica (Quigley y Bahmueller 1991).

Directrices estatales del currículo

Las directrices estatales del currículo ofrecen otra valiosa fuente de información. California está a la cabeza en el desarrollo de unas directrices de ciencias histórico-sociales, que incluye historia, geografía, economía, ciencias políticas, antropología, psicología, sociología y humanidades (California State Department of Education 1988). Las directrices incluyen tres áreas de objetivos principales. Cada área contiene líneas curriculares que suben en espiral a lo largo de la educación de un alumno:

- Objetivos de conocimiento y comprensión cultural
  - — familiarización con la historia
  - — familiarización con la ética
  - — familiarización con la cultura
  - — familiarización con la geografía
  - — familiarización con la economía
  - — familiarización con la sociopolítica

- Objetivos de aprendizaje de destrezas y participación social
  - — destrezas básicas de estudio
  - — destrezas de pensamiento crítico
  - — destrezas de participación
- Objetivos de comprensión democrática y valores cívicos
  - — identidad nacional
  - — patrimonio constitucional
  - — valores cívicos, derechos y responsabilidades

Cada una de estas áreas comprende un número de objetivos de aprendizaje que podrían ser temas de evaluación. Por ejemplo, las destrezas de participación en el apartado "objetivos de aprendizaje de destrezas y participación social" incluyen destrezas personales, destrezas de interacción en grupo y destrezas de participación social y política. "La familiarización con la economía" incluye cuestiones específicas relacionadas con los problemas económicos fundamentales a los que se tiene que enfrentar cada sociedad; sistemas económicos comparativos; objetivos económicos fundamentales, rendimiento y problemas de la sociedad; y el sistema económico internacional.

Connecticut ha formulado el "Common Core of Learning" (1987), una serie de criterios de aprendizaje comunes dirigido a alumnos de educación media. Los criterios incluyen *destrezas genéricas*, que se encuentran en todas las disciplinas, y las *grandes ideas* y *destrezas, conceptos, procesos* y *técnicas* que caracterizan una disciplina específica. Estas destrezas genéricas constituyen un punto de partida para establecer las finalidades clave de los alumnos en cualquier disciplina. Estas destrezas genéricas son:

- comunicarse con claridad;
- formular preguntas;
- formular problemas;
- pensar y razonar;
- resolver problemas complejos;
- sintetizar conocimientos obtenidos de varias fuentes; y
- cooperar y colaborar.

Las destrezas, las técnicas y los procesos científicos de Connecticut, que también son genéricas, incluyen:

- desarrollar una hipótesis;
- diseñar experimentos;
- sacar conclusiones a partir de unos datos;
- observar y analizar semejanzas y diferencias entre los diversos fenómenos; y
- trabajar con material de laboratorio.

Otras fuentes

Otra fuente de información para el desarrollo de la evaluación la encontramos en directrices que se han hecho para evaluaciones nacionales e internacionales. La "Evaluación Nacional del Progreso Educativo" ("National Assesment of Educational Progress"; NAEP) evalúa regularmente el rendimiento escolar en asignaturas como matemáticas, lengua, ciencias, historia, geografía y en el alfabetismo adulto. Como parte de su operación, NAEP lleva a cabo un proceso de consenso nacional que define las directrices de contenido de cada evaluación y establece una serie prioridades para el éxito escolar. (Para más información sírvase comunicarse con el Educational Testing Service, Rosedale Road, Princeton, NJ 08541; teléfono (609) 921-9000).

Aprovechar las campañas de reestructuración escolar

Los grupos involucrados en campañas de reestructuración escolar son una fuente adicional. Por ejemplo, un aspecto fundamental en la campaña de la "Coalition of Essential Schools" consiste en una exposición final en la que los alumnos demuestran sus habilidades. Los miembros de esta coalición han estudiado con mucho detenimiento cuáles deberían ser estas habilidades. Diversas escuelas han propuesto perfiles de las habilidades que sus alumnos deben haber adquirido al finalizar el año lectivo o al graduarse de la escuela. A continuación presentamos algunos ejemplos:

- Los alumnos de esta asignatura podrán comprender mejor muchas de las cuestiones a las cuales se enfrenta su generación. Serán capaces de hablar y escribir sobre temas actuales con conocimiento, curiosidad y honestidad. Y reflexionarán detenidamente sobre el papel que desempeñan en calidad de presentadores de información (Parkway South, temas contemporáneos).
- Los alumnos...principalmente aprenderán a aplicar los conceptos geométricos a situaciones del mundo real (Sullivan High School, matemáticas).
- Los alumnos en esta asignatura aprenderán a trabajar en equipo para producir trabajo informativo de gran calidad. Obtendrán conocimientos sólidos de las técnicas de campo necesarias para estudiar ecología. Se sentirán orgullosos al saber que han contribuido de forma tangible a su comunidad...Y quizás, lo más importante, obtendrán una buena comprensión y un sentimiento de responsabilidad por el entorno natural en el que viven (Sullivan High School, ecología).
- Al finalizar sus estudios, los alumnos de esta escuela sabrán cómo explorar las ideas a fondo y de manera significativa y serán capaces de expresar sus pensamientos de manera elocuente, coherente y correcta (Sullivan High School, humanities).

- Los alumnos al finalizar sus estudios en el Metro High School tendrán una clara idea de cuáles son sus intereses y aptitudes individuales. Al finalizar la escuela se sentirán seguros de tener las destrezas necesarias para alcanzar sus metas, las cuales han estudiado y planeado con mucho detenimiento.
- Al finalizar sus estudios, los alumnos de esta escuela serán adultos motivados, perspicaces y exigentes que piensan de una forma independiente y responsable. Tendrán unos conocimientos amplios del temario, unas destrezas de aprendizaje bien desarrolladas...(Crefeld School).

Declaraciones de este tipo y descripciones de cómo se llevan a cabo las exposiciones se encuentran en *The Exhibitions Collections*, diseñadas y distribuidas por la Coalición de la Universidad de Brown. (Sírvase comunicarse con Joe McDonald, Coalition of Essential Schools, Brown University, Box 1969, Providence, RI 02912; teléfono (401) 863-3384; FAX (401) 863-2045.)

Entre otras fuentes de objetivos escolares significativos e innovadores se encuentran el Proyecto de Aprendizaje Acelerado de Henry Levin (Henry Levin's Accelerated Learning Project) (1989); el Proyecto de James Comer (Comer y Haynes-Norris 1991); el Proyecto Foxfire de Elliot Wiggington (Puckett 1989); y el currículo de Otras Maneras de Aprender del Instituto Galef (Galef Institute's Different Ways of Knowing curriculum) (Galef Institute 1992).

Considerar objetivos interdisciplinarios

Muchas de las nuevas directrices que están siendo desarrolladas muestran un creciente reconocimiento de los objetivos interdisciplinarios. Los criterios de matemáticas de la NCTM ponen atención en las destrezas comunicativas. La AAAS considera que las matemáticas, las ciencias y la tecnología están integralmente relacionadas y recomienda que los alumnos entiendan cómo fue que las ideas importantes de la ciencia surgieron de sus particulares contextos históricos, culturales e intelectuales. En las directrices de las ciencias historicosociales de California y en muchas de las exposiciones de la Coalición de Escuelas se describe un método curricular interdisciplinario. Durante el desarrollo de la evaluación, también se puede considerar la posibilidad de destacar los objetivos interdisciplinarios para los alumnos.

Consultar con los colegas

¿Cuáles son los objetivos específicos del currículo de nuestra clase y de nuestro programa docente? A la hora de consultar fuentes disponibles que nos respondan a esta pregunta no debemos dejar a un lado a nuestros colegas. La colaboración entre colegas fomenta llegar a un consenso a nivel de escuelas y a mejorar las

evaluaciones. Si trabajamos en la evaluación de un departamento, de una escuela o de un distrito, quizás queramos incluir a padres, miembros de la comunidad y representantes del mundo de los negocios en el proceso de preparación.

Establecer prioridades significativas: una propuesta difícil

Ya estemos solos o formemos parte de un grupo, probablemente encontraremos que ahora tenemos en nuestro haber una larga lista de posibles objetivos para evaluar el rendimiento. Conforme revisamos esta lista, ya sea solos o en colaboración con otros, podemos pensar en las siguientes preguntas para lograr enfocarnos en la evaluación:

1. ¿Cuánto **tiempo** les llevará a los alumnos desarrollar o adquirir la destreza o habilidad? Si la respuesta es una hora, un día o una semana, probablemente no valga la pena invertir el tiempo y el esfuerzo que requeriría hacer una evaluación completa.
2. ¿Cómo se relaciona la destreza o habilidad deseada con **otras destrezas cognitivas, sociales y afectivas complejas**? Se debería dar una mayor prioridad a las destrezas que están integralmente relacionadas con otras destrezas importantes. Es necesario dar prioridad a aquéllas que puedan aplicarse a varias situaciones.
3. ¿Cómo se **relaciona** la destreza o habilidad deseada con **los objetivos escolares y curriculares a largo plazo**? Hay que dar prioridad a los objetivos a largo plazo o a los componentes integrales de objetivos importantes a largo plazo.
4. ¿Cómo se **relaciona** la destreza o habilidad deseada con **los planes de mejora de la escuela**? Es necesario dar prioridad a aquéllas que tienen más peso dentro del plan.
5. ¿Cuál es la **importancia intrínseca** de las destrezas o habilidades deseadas? Evidentemente hay que dar prioridad a aquéllas que son importantes y descartar aquéllas que no son más que objetivos triviales o superficiales. (Aunque parezca obvio, pensemos en todas esas preguntas sobre temas sin importancia a las cuales hemos respondido en los exámenes.)
6. ¿Se pueden **enseñar y son posibles de alcanzar** las destrezas y habilidades deseadas **para nuestros alumnos**? A la vez que intentamos desafiar a los alumnos y sacar a relucir lo mejor de ellos, tenemos que tomar en cuenta si los alumnos poseen las necesarias destrezas, conceptos y conocimientos básicos que son fundamentales para alcanzar los objetivos. También debemos considerar si tenemos el material y aptitud necesaria para ayudarles a alcanzar dichos objetivos.

Como resultado de ese tipo de proceso para tomar decisiones, se puede identificar una serie imprescindible de destrezas y habilidades. Se debe describir cada una de ellas con suficiente detalle para que otros entiendan su significado. Aunque se necesite revisar estas primeras descripciones, esta lista de prioridades marcará los objetivos principales para el diseño de una evaluación.

Para aprender cómo desarrollar y llevar a cabo evaluaciones alternativas, se puede empezar con una evaluación en concreto. Hay que tomar en consideración los objetivos escolares que tengan más peso, el período del año y la parte del currículo donde nos encontremos. Luego se debe designar una de las finalidades prioritarias como objetivo principal. El siguiente paso será identificar las tareas apropiadas para la evaluación de este objetivo.

Referencias bibliográficas

American Association for the Advancement of Science. (1989). *Project 2061: Science for all Americans*. (Publication No. 89-01S). Washington, D.C.: AAAS.

California State Department of Education. (1988). History/ *Social Science Framework for California Public Schools: Kindergarten Through Grade 12*. Sacramento, Calif.: California State Department of Education.

Caterall, J. (1991). *Different Ways of Knowing*. Los Angeles: Galef Institute.

Comer, J., and M. Haynes-Norris (January 1991). "Parent Involvement in Schools: An Ecological Approach." *Elementary School Journal* 91, 3: 271-277.

Connecticut State Board of Education. (January 1987). *Connecticut's Common Core of Learning*. Hartford, Conn.: Connecticut State Department of Education.

Galef Institute. (1992). *Different Ways of Knowing*. (Information packet.) Los Angeles, Galef Institute (11150 Santa Monica Boulevard, Los Angeles, CA 90025).

Levin, H. (May 1989). *Accelerated Schools: A New Strategy for At-risk Students* (Policy Bulletin No.6). Bloomington, Ind.: Consortium on Educational Policy Studies.

National Council of Teachers of Mathematics. (March 1989). *Curriculum and Evaluation Standards for School Mathematics*. Reston, Va.: NCTM.

Puckett, J.L. (1989). "Who Wrote Foxfire? A Consideration of Ethnohistorical Method." *Journal of Research and Development in Education* 22, 3: 71-78.

Quigley, C.N., and C.F. Bahmueller. (1991). *CIVITAS: A Framework for Civic Education*. Calabasas, Calif.: Center for Civic Education.

4

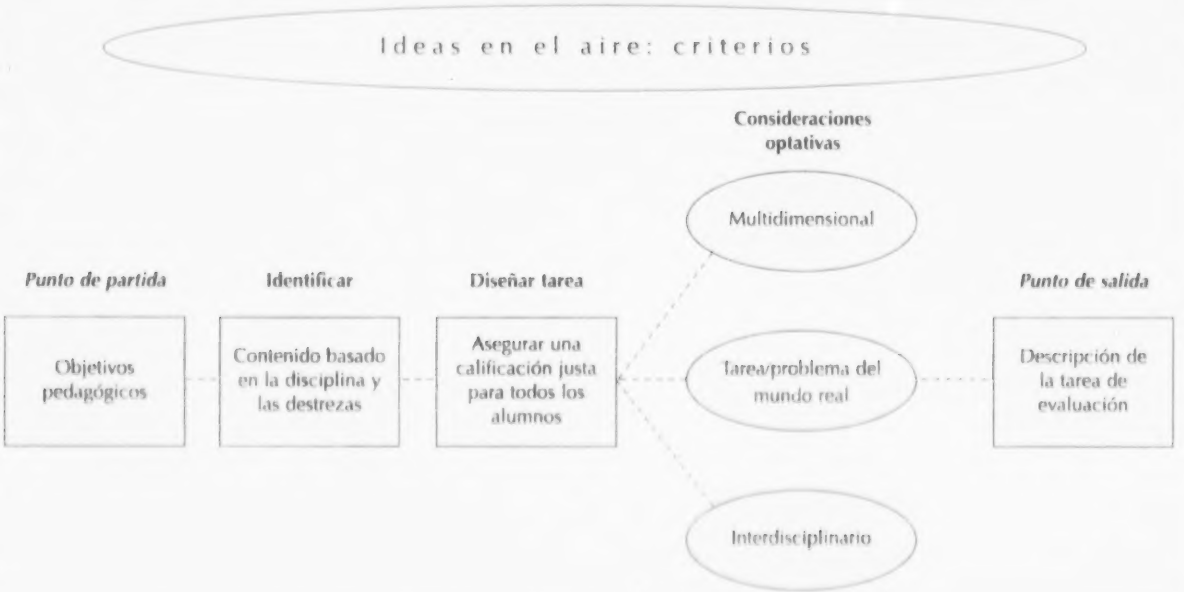
Selección de tareas de evaluación

La clave de una buena evaluación está en ajustar la tarea de evaluación a los resultados que se desean de los alumnos (los conocimientos, destrezas y disposiciones que se identificaron en la primera planificación de la evaluación). ¿Qué tareas o trabajos representan estas habilidades que se buscan? Se podrían crear muchas posibilidades interesantes y apropiadas. Al considerar las tareas de evaluación, la mejor opción sería elegir aquéllas que se creen más cerca de los objetivos de la enseñanza y que permiten a los alumnos demostrar su progreso y capacidad.

Al intentar formular tareas interesantes para los alumnos, es posible que algunas no correspondan con las prioridades que en un principio se tenían, pero sí podrían representar objetivos importantes que se dejen de lado. Esto es un ejemplo de cómo el proceso de desarrollo de la evaluación no es lineal. Las decisiones que se toman a cada paso se ven influidas por las que les preceden y las que les siguen. Hay muchos maestros que encuentran más fácil describir cuáles son los objetivos válidos para los alumnos *después* de pensar en los tipos de trabajos que consideran más interesantes, desafiantes y útiles para ellos.

Hay que tener en cuenta algunas cuestiones a la hora de diseñar tareas de evaluación que sean apropiadas. La ilustración 4.1 nos muestra una perspectiva conceptual de algunas de estas cuestiones. En esta ilustración se aprecia claramente la dificultad de pensar en tareas de evaluación sin tener en cuenta a la vez los

Ilustración 4.1.
Mapa de tareas de evaluación



BEST COPY AVAILABLE

criterios que se van a utilizar para evaluar el rendimiento en esas tareas. Aunque en el capítulo 5 tratamos los criterios del rendimiento, veremos cómo esta separación demuestra que llevar a cabo la planificación de una evaluación no es tarea ni fácil ni lineal.

Elegir las tareas convenientes

La respuesta a las siguientes preguntas nos ayudará a seleccionar las tareas de evaluación.

¿Hay concordancia entre la tarea y las intenciones didácticas específicas?

Al intentar evaluar un único aspecto, es fácil que surjan ideas de posibles tareas. Por ejemplo, si se quiere que los alumnos se comuniquen con eficacia, parece obvio que se les debe pedir que escriban algo. Pero ¿qué deben escribir? Si todavía no se ha marcado objetivos específicos didácticos—por ejemplo, los tipos de textos escritos que se requiere de los alumnos: narrativos, expositivos y persuasivos—ahora es el momento de hacerlo. De igual manera, si se quiere que los alumnos sean capaces de aplicar el método científico, parece lógico pedirles que lleven a cabo experimentos o realicen estudios enfocados a algún tema, pero también es necesario decidir qué contenidos y destrezas específicas debería incluir la tarea. ¿Qué tipo de experimentos? ¿Qué clase de estudios: un estudio de la composición del abono? ¿Una investigación sobre las necesidades de la comunidad? ¿Un estudio escolar de los hábitos alimenticios? Es importante que la tarea de la evaluación concuerde con la finalidad didáctica que se pretende evaluar.

¿Es adecuada la tarea al contenido y destrezas que se espera que los alumnos aprendan?

Según las teorías modernas de aprendizaje, el contenido y el proceso están íntimamente ligados. Por ejemplo, el proceso de razonamiento a seguir en las ciencias sociales es diferente al de las matemáticas. Para poder resumir por escrito un contenido biológico se requieren conocimientos y destrezas distintas a aquéllas que se necesitan para hacer un resumen de un texto literario. Por ello, además de especificar la naturaleza general de la tarea, se necesitará pensar sobre los temas específicos o áreas temáticas que se van a pedir a los alumnos. Por ejemplo, si se desea que los alumnos escriban textos persuasivos, ¿cuál sería el tema de su trabajo? ¿Sería un problema hipotético, un problema escolar, un dilema personal, un tema de actualidad, un asunto local, una solución matemática o un problema

ético? ¿Qué campo de contenido se espera que utilicen—sus conocimientos previos, una investigación adicional o sus conocimientos personales?

Supongamos que se quiere que los alumnos sean capaces de realizar experimentos científicos, por ejemplo experimentos químicos, para la resolución de problemas. Al decidir la tarea de evaluación, habrá que tener en cuenta otros temas de contenido específicos. ¿Con qué tipos de sustancias deberían trabajar? ¿Qué clase de problemas—análisis, diseño o evaluación? ¿Y qué tipo de propiedades químicas se quiere que incorporen? ¿Qué tipo de material deberían saber manejar? En resumen, ¿cuál es el campo de contenidos, conceptos, principios y técnicas con las que los alumnos deberían estar familiarizados? Y basándose en éstos ¿cuál sería un buen ejemplo de lo que se espera de los alumnos? ¿Se quiere que analicen sustancias desconocidas con determinadas propiedades, que predigan qué producto le iría mejor a un determinado propósito, o que determinen qué cultivo sería el más eficaz y económico para acabar con el hambre?

¿Permite la tarea que los alumnos demuestren su progreso y capacidad?

Al pensar en el contenido específico que se espera con relación al rendimiento de un alumno, surgen una serie de temas interrelacionados sobre la imparcialidad de la tarea y la posible falta de objetividad. ¿Qué conocimientos previos del alumnos presupone la tarea? ¿Han tenido los alumnos la oportunidad de adquirir estos conocimientos? ¿Incluye la tarea destrezas que son relevantes para el deseado objetivo de la evaluación? En otras palabras, ¿es la tarea una evaluación justa de lo que saben los alumnos y podrán los alumnos demostrar sus aptitudes y capacidades? Tomando otro ejemplo de la expresión escrita, es sabido que los alumnos necesitan de conocimientos previos para los temas sobre los cuales tienen que escribir. Sin estos conocimientos, no podrían decir nada. La estimación del nivel de las destrezas de expresión escrita siempre está unida a lo que los alumnos saben (o no saben) sobre el tema en cuestión. A la vez que se formulan temas específicos para los alumnos, hay que prestar atención a la interrelación que existe entre el contenido y la destreza. Hay que evitar que las habilidades de los alumnos para demostrar sus destrezas se vean entorpecidas por la inclusión en la evaluación de algo que puede ser irrelevante a los objetivos. Por ejemplo, si los alumnos no están al corriente de los temas de actualidad, no se puede esperar que escriban un texto elocuente y tomen una postura sobre una cuestión nacional del momento. O si los alumnos no son buenos lectores no se debe entorpecer su capacidad para demostrar sus destrezas en la expresión escrita pidiéndoles, por ejemplo, que escriban sobre un artículo de *The New York Times*. Por supuesto se debe tener objetivos de lectura para los alumnos y quizá se quiera que adquieran conocimientos sobre temas de actualidad, pero hay que evitar frustrar accidentalmente su capacidad para demostrar destrezas específicas, o etiquetarles equivocadamente como no aptos, apoyándose en una tarea inapropiada o en la falta de oportunidad para la adquisición de los necesarios conocimientos previos y destrezas.

Una solución al dilema de los conocimientos previos es facilitar a los alumnos el acceso a fuentes relevantes, que saben manejar, como parte del proceso evaluativo. Por ejemplo, en Connecticut los alumnos de educación media dentro de la asignatura de química diseñan y llevan a cabo experimentos en los que tienen que distinguir sustancias desconocidas y averiguar cuál gaseosa contiene azúcar y cuál no. La tarea evalúa aspectos diferentes dependiendo de los libros de texto y de otras fuentes que se permite consultar a los alumnos. Si los maestros limitan tales fuentes, el rendimiento del alumno dependerá de si recuerda las pruebas específicas para los azúcares y su composición química. Aquellos alumnos que no recuerdan con facilidad estos hechos, no llegarán muy lejos en la preparación o en la realización de las pruebas apropiadas. Por otro lado, si los maestros permiten que los alumnos tengan acceso a las fuentes relevantes, la tarea evalúa más directamente si los alumnos saben cómo diseñar y llevar a cabo experimentos científicos, suponiendo, por supuesto, que sus libros de texto no contengan la solución del problema. ¿Cuál es el mejor método? La respuesta depende de las intenciones y expectativas del profesor.

Otra solución al dilema de los conocimientos previos es proporcionar a los alumnos una variedad de opciones en la tarea evaluativa, por ejemplo, dándoles libertad para escoger la forma de expresión que deseen—expresión escrita, oral, visual o musical, y una variedad de tareas de dificultad diversa.

¿Utiliza la evaluación tareas auténticas del mundo real?

Los teóricos contemporáneos del currículo destacan la importancia de involucrar a los alumnos en tareas auténticas y del mundo real ya que parecen más motivadoras y poseen una mayor transferibilidad que las tareas académicas más tradicionales y descontextualizadas. También proponen estos teóricos que involucrar a los alumnos en *el proceso* de una disciplina mientras adquieren o demuestran conocimientos en esa disciplina es una potente estrategia de aprendizaje. Por ejemplo, la tarea de química de Connecticut trata a los alumnos como científicos y les pregunta sobre algo con lo que están muy familiarizados en el mundo real.

De igual manera el "Prototipo de Evaluación de Contenido" (Content Assesment Prototype) en historia, desarrollado por Eva Baker y colegas (1992) en CRESST, trata a los alumnos como historiadores en tareas que son reales. A los alumnos se les pide que lean el material de fuentes primarias, por ejemplo una versión abreviada de los debates Lincoln-Douglas. Luego tienen que recurrir a sus conocimientos previos y a lo que conocen del tema para explicar los hechos históricos tratados en estos documentos e incorporar el contenido histórico—los problemas y temas a los que se enfrentaba la nación antes de que se desatara la Guerra Civil. Con el fin de darle un objetivo auténtico a la tarea, el protocolo de la evaluación también establece un público apropiado al que se dirigirá las respuestas de los alumnos.

Los problemas del mundo real, las técnicas realistas y los auténticos públicos lectores proporcionan innumerables posibilidades para las tareas. Los maestros de ciencias sociales, por ejemplo, pueden pedir a los alumnos que investiguen un

problema de actualidad y que luego escriban una carta al Congreso de la Unión o al ayuntamiento, o que diseñen un anuncio de interés público en el que se recomiende una solución. Los maestros de ciencias pueden invitar a los alumnos a que escriban cartas a los periódicos o a los senadores de su estado, o que filmen un video sobre problemas ecológicos. Los maestros de matemáticas pueden alentar a los alumnos para que lleven a cabo una investigación sobre las necesidades de la comunidad y que redacten un informe o calculen cuánto dinero necesitarían para llevar a cabo uno de sus objetivos futuros como, por ejemplo, la compra de un coche, teniendo en cuenta el precio, los gastos del préstamo/intereses, el seguro, los impuestos, el permiso de circulación, el mantenimiento, la gasolina, etcétera.

¿Se presta la tarea a un enfoque interdisciplinario?

Los problemas auténticos y del mundo real no siempre se ajustan perfectamente a los distintos dominios curriculares. Más bien, los alumnos tienen que hacer uso de conocimientos de varias disciplinas y perspectivas. La "carta al editor proponiendo una solución a un problema ecológico" implica las destrezas comunicativas de los alumnos, sus destrezas científicas para comprender problemas ecológicos específicos y sus destrezas interpersonales sabiendo quién es su público. En otro caso, un proyecto de investigación podría requerir de un alumno que investigara un tema, diseñara un estudio empírico basado en los datos y principios científicos que investiga, utilizara destrezas matemáticas para analizar y mostrar los datos de su estudio y que aplicara tanto sus destrezas científicas como las comunicativas para resumir los resultados y comunicárselos a otros.

Las tareas interdisciplinares ofrecen además otras ventajas como es el factor tiempo y una más acertada puntuación. En realidad, tareas que implican un rendimiento significativo muchas veces llevan largos períodos de tiempo y puede que simplemente no haya suficiente tiempo para evaluar las áreas de contenido por separado. Las tareas interdisciplinares ayudan a los maestros a evitar este posible problema.

¿Se puede estructurar la tarea para evaluar distintos objetivos?

Es evidente que las tareas interdisciplinares podrían evaluarse desde las diferentes perspectivas de las disciplinas implicadas. Por ejemplo, puesto que la carta al editor requiere destrezas de expresión escrita, destrezas interpersonales y conocimientos científicos, se podría calificar el rendimiento en cada una de estás áreas por separado.

La mayoría de las tareas de evaluación diseñadas para evaluar objetivos significativos también incorporan una serie de destrezas cognitivas, metacognitivas, afectivas y sociales. Por ejemplo, el ejercicio de química de las gaseosas, que se realiza durante varios días, incluye los siguientes componentes: trabajo en equipo, trabajo individual, un informe oral y la reflexión individual y de

equipo. En grupos reducidos, los alumnos deben en primer lugar participar en la técnica de lluvia de ideas para obtener una lista de posibles pruebas que les permita averiguar cuál de las dos gaseosas es la que contiene azúcar. Luego realizan dos pruebas, analizan sus resultados y presentan un informe oral a la clase. También se pide a cada alumno que resuelva otro problema de análisis químico parecido. Los alumnos reflexionan sobre los puntos fuertes y débiles de su rendimiento a nivel individual y a nivel de grupo, sobre el rendimiento de otros miembros del grupo y sobre sus actitudes frente a la tarea.

La estructuración de "megatareas" que evalúen distintas finalidades requiere de mucho ingenio. Si los objetivos de mayor prioridad abarcan el trabajo en equipo y el trabajo individual, se puede invitar a los alumnos a trabajar en equipo para resolver un problema, pero habrá que motivarlos para que trabajen de forma individual durante una o más etapas del proyecto, poniendo a cada alumno individualmente a recopilar y resumir información para un proyecto de grupo. Alternativamente, si se desea que los alumnos trabajen en equipo para definir y resolver un problema en particular, pero cada alumno deberá presentar un informe de lo que el grupo ha descubierto. Si se quiere medir hasta qué punto los alumnos aceptan los desafíos e intentan resolver los problemas a pesar de los esfuerzos y dificultades que conlleva, habría que incluir suficientes elementos desafiantes y de elección propia en la tarea de evaluación para que los alumnos puedan mostrar más o menos entusiasmo, esfuerzo y empeño; e incluir para el evaluador alguna manera de observar la conducta y el afecto. En el capítulo 5 se discuten los criterios con los que se pueden valorar la conducta y el afecto.

Hay que tener en cuenta que aunque es provechoso y eficaz diseñar estas tareas de evaluación multidimensionales, complejas y ricas, también tiene sus desventajas. La más importante entre éstas es obtener de las respuestas de los alumnos aquello que se puede atribuir a la destreza que han adquirido, aquello que representa conocimientos previos e incluso la determinación del nivel de logro de cada alumno. Por ejemplo, los alumnos con destrezas de expresión escrita limitadas no serán capaces de demostrar adecuadamente su nivel real de comprensión por medio de la escritura. Los alumnos que no están muy motivados quizás desistan de una tarea larga antes de poder demostrar su nivel de competencia. Y si los alumnos participan en trabajos en equipo como parte de la tarea, es posible que sea más difícil evaluar los logros de cada individuo.

Buscar ideas buenas para las tareas

La técnica de lluvia de ideas con los colegas es una estrategia buena para que surjan las primeras ideas para diseñar buenas tareas de evaluación. Se puede empezar pensando en los proyectos docentes más complejos y exitosos que se han llevado a cabo en el pasado. Hay que recordar la primera regla de la lluvia de ideas: ser creativo, anotar todo lo que nos venga a la mente y no criticar ninguna idea hasta que estén todas sobre la mesa. Luego se puede combinar, refinar y mejorar los mejores aspectos de cada idea que haya surgido.

Además de hacer uso de las ideas propias, hay que aprovechar los esfuerzos de los demás. Se puede adaptar y mejorar las ideas que se han obtenido de revistas profesionales, congresos y cursos de formación, de observaciones de las clases de otros maestros, etcétera. Hay que tener en cuenta que un gran número de estados, distritos escolares y escuelas están trabajando para desarrollar estos nuevos métodos de evaluación. Si un estado ya tiene su propia evaluación, ésta puede ser una fuente de ideas. CRESST está armando una base de datos de los esfuerzos que se han hecho a nivel nacional, los cuales distribuirán a través de ERIC. Esta base de datos incluirá muestras de evaluaciones de rendimiento en una amplia gama de asignaturas de distintos cursos. Aunque ninguna de estas muestras se ajuste totalmente a las necesidades y objetivos de cada uno, se puede tomar prestado las ideas sobre la evaluación que representan, las escalas que utilizan para puntuar el rendimiento de un alumno, etcétera. Incluso aunque no seamos maestros de química, nos podría intrigar el método de Connecticut para evaluar el trabajo en equipo y podríamos adaptar sus escalas a nuestro propio trabajo en equipo. La evaluación Lincoln-Douglas/Guerra Civil de comprensión y explicación, descrita anteriormente, puede sugerirnos un método parecido para la evaluación de alumnos en asignaturas como son las ciencias sociales, naturales o crítica del arte.

Si las evaluaciones que se están desarrollando son parte del esfuerzo de toda una escuela, hay que tener en cuenta a otros miembros de la comunidad escolar—padres, representantes de negocios y miembros de la comunidad. Los individuos que no forman parte de la escuela pueden ser de gran ayuda a la hora de concebir tareas auténticas del mundo real que demuestren destrezas importantes de razonar, de resolver problemas y de comunicación. También pueden ayudar en calidad de "revisores de tareas" y para advertirnos sobre los tipos de conocimientos, relevantes e irrelevantes, que estas tareas representan.

Describir la tarea de evaluación

Hay que especificar o documentar cuidadosamente las tareas de evaluación formales para que otros puedan interpretar los resultados o repetir los métodos con otros alumnos en otras situaciones. O incluso, lo que es más importante, ya que se supone que las evaluaciones representan el rendimiento de un alumno en un dominio mayor, es imprescindible que se sepa cuál es ese dominio más amplio. Una descripción de la tarea ayuda a definir el dominio mayor, proporciona los cimientos para otras evaluaciones específicas que puedan hacerse y permite revisar el trabajo y localizar problemas importantes antes de someterlas a prueba con los alumnos.

Aunque la naturaleza de la tarea de evaluación dictará lo que se necesita especificar, normalmente se requiere especificar los siguientes aspectos:

- ¿Cuál es el fin(es) que se pretende con la evaluación?
- ¿Cuáles son los contenidos/temas que pueden entrar?
- ¿Cuál es la naturaleza y el formato de las preguntas que se hacen a los alumnos? ¿A quién van dirigidas las respuestas?

- ¿Es un trabajo individual o de equipo? Si es trabajo de equipo, ¿qué roles se han de jugar?
- ¿Qué opciones/elecciones se permiten? ¿Cuáles son las opciones de respuesta? ¿Qué incluyen, por ejemplo carpetas de trabajo? ¿Quién hace las selecciones—el maestro o los opciones de respuesta? ¿Qué incluyen, por ejemplo, las carpetas de trabajo? ¿Quién hace la selección, el maestro, los alumnos, o ambos?
- ¿Qué material/equipo/fuentes tendrán los alumnos a su disposición? ¿Hay algunas especificaciones?
- ¿Qué instrucciones se da a los alumnos?
- ¿Qué restricciones administrativas hay? ¿Cuánto tiempo tienen los alumnos? ¿Cuál es el orden de la tarea? ¿Cómo se responderá a las preguntas de los alumnos? ¿Qué tipo de ayuda se permitirá?
- ¿Qué baremo y procedimiento de puntuación se va a seguir?

La ilustración 4.2 nos proporciona un ejemplo de plantilla para la descripción de la tarea. La lista de control resume tanto los asuntos más importantes relacionados con la creación de las tareas de evaluación como los asuntos de puntuación que hay que establecer y que trataremos en el capítulo cinco.

Asegurar que las tareas conducen a evaluaciones fiables

Dada la complejidad del desarrollo de una tarea, habrá que revisar las tareas antes de llevarlas a cabo con los alumnos. Estos criterios pueden ayudar a formar una crítica sobre las ideas de evaluación antes de desarrollarlas completamente:

- ¿**Se ajustan las tareas a los objetivos importantes** que se han establecido para los alumnos? ¿Reflejan estos objetivos destrezas complejas de razonamiento, como el análisis y la síntesis?
- ¿Constituyen un tipo de problema **perdurable**—los tipos de problemas y situaciones a los que los alumnos probablemente tengan que enfrentarse una y otra vez en la escuela y en el futuro?
- ¿Son las tareas **justas y objetivas**? Por ejemplo, ¿favorecen a los chicos o a las chicas, a los alumnos que han vivido en una región o lugar en particular, a los alumnos que tienen una herencia cultural particular, o aquéllos a quienes sus padres tienen los medios para comprar determinado material?
- ¿Tendrán **crédito** las tareas para los sectores importantes? ¿Serán vistas como significativas y desafiantes por parte de alumnos, padres y maestros? ¿Dependen las tareas de un contenido curricular de calidad?
- ¿Serán las tareas lo suficientemente **significativas** y atractivas como para que los alumnos se sientan motivados a mostrar sus capacidades? ¿Incluyen las tareas problemas, situaciones y públicos reales?

Ilustración 4.2
Lista de control para la descripción de la tarea

Objetivos que se han de evaluar	• Descripción de objetivos didácticos • Contenido/temas pertinentes • Reglas/Proceso de selección
Proceso de administración de la evaluación	• Roles individuales/de grupo • Materiales/equipo • Instrucciones de la administración • Ayuda permitida • Tiempo permitido
Pregunta/problema/estímulo reales	• Formato • Público • Opciones disponibles • Instrucciones para alumnos
Puntuación	• Rúbrica • Procedimientos de puntuación • Utilización de las calificaciones

- ¿Tienen las tareas **relación con la enseñanza** o se pueden siquiera enseñar? ¿Representan las tareas las destrezas y conocimientos que los alumnos pueden adquirir y de los que se tiene el material y la pericia adecuada para enseñarlos?
- ¿Son **viables** las tareas para llevarse a cabo en la clase o en la escuela en términos de espacio, equipo, tiempo, dinero, etcétera? ¿Pueden los alumnos llevarlas a cabo junto con sus obligaciones extraescolares, que incluye a la familia y otras cosas que requieren de tiempo, su acceso a bibliotecas y otras fuentes, y asequibilidad.

Estos criterios se han derivado de los criterios más generales de CRESST para asesorar la calidad de la evaluación (Linn et al. 1991). Consultarlos ayuda a asegurar que las evaluaciones den lugar a inferencias válidas sobre los alumnos y los programas.

Referencias bibliográficas

Baker, E.L., P.R. Aschbacher, D. Niemi y E. Sato. (1992). *CRESST Performance Assessment Models: Assessing Content Area Explanations*. Los Angeles: University of California, Center for Research on Evaluation, Standards and Student Testing.

Linn, R.L., E.L. Baker y S.B. Dunbar. (1991). "Complex, Performance-based Assessment: Expectations and Validation Criteria." *Education Researcher* 20, 8: 15-23.

5

Establecer criterios

Los criterios que se utilizan para evaluar el rendimiento del alumno constituyen la base de la evaluación alternativa. Aunque hemos analizado la selección y la descripción de las tareas de evaluación por un lado, y el establecimiento de los criterios de calificación por otro, es necesario tener en cuenta que estos tres aspectos de la evaluación están íntimamente relacionados. En ausencia de criterios, las tareas de evaluación son simplemente eso, tareas o actividades educativas. Quizá lo más importante es que los criterios de calificación hacen público aquello que se está evaluando y, en muchos casos, los parámetros de un rendimiento satisfactorio. Por consiguiente, los criterios comunican los objetivos y los niveles que se han de alcanzar.

Al igual que la propia "evaluación alternativa", también los criterios para evaluar el rendimiento del alumno han recibido varios nombres, entre ellos criterios de calificación, directrices para la calificación, rúbricas y rúbricas de calificación. Para nuestro propósito, entendemos todos estos términos como una **descripción de las dimensiones** que se utilizan para evaluar el rendimiento del alumno, un **baremo de valores** para calificar esas dimensiones y, cuando sea apropiado, los **estándares** prefijados en la evaluación del rendimiento.

Tomemos un ejemplo común de las ciencias sociales. Se pide a los alumnos que hagan una presentación en grupo y que individualmente redacten informes para evaluar su nivel de comprensión de historia. Puesto que se pretende evaluar tres destrezas—destrezas orales, escritas y de trabajo en grupo en relación con la

historia—hay que tener en cuenta criterios de calificación para cada destreza. El cuadro 5.1 en las páginas 46–47 muestra ejemplos de criterios de calificación para sólo una de estas destrezas, una evaluación de un trabajo de historia hecho en grupo diseñado por el Programa de Evaluación de California (California Assessment Program).[1]

El ejercicio de procesar información en grupo explota cuatro **objetivos de aprendizaje:** el aprendizaje en grupo, el razonamiento crítico, la comunicación y los conocimientos de historia. Para cada objetivo se especifican las **dimensiones** de calificación y los niveles de rendimiento que se diferencian en un **baremo de valoración**. Finalmente, la guía de calificación incluye una **evaluación** de cada nivel de rendimiento, que describe el rendimiento no sólo en términos de logros, sino también en el éxito de éstos, en una escala que va de insuficiente a sobresaliente.

Entender la necesidad de los criterios

Los criterios son necesarios porque nos ayudan a valorar de manera fiable, justa y válida el complejo rendimiento humano. Los criterios de calificación guían las valoraciones y hacen público las bases de estas valoraciones para los alumnos, padres y otros. Calificar un examen tipo test no requiere de una valoración complicada; sin embargo, el juicio humano sigue siendo un factor importante ya que el responsable de diseñar el examen formula las preguntas y decide lo que constituye las mejores respuestas. Para la persona que corrige el examen, un alumno selecciona o no la respuesta correcta; no se necesita juicio alguno. Cuando utilizamos exámenes con respuestas que se han de seleccionar, realmente estamos corroborando juicios sobre lo que se considera un rendimiento apto que están implicados en la clave de respuestas. Por consiguiente toda evaluación, ya sea por medio de exámenes donde se tiene que seleccionar o formular respuestas, contiene un elemento subjetivo o de juicio humano.

Las evaluaciones alternativas invitan a elegir entre una gama más amplia de posibles respuestas. En lugar de calificar las respuestas como bien o mal, las evaluaciones alternativas valoran la capacidad de llegar a una respuesta compleja y a veces hasta el proceso para llegar a ella. Para hacer tales valoraciones y para asegurar su validez, constancia e imparcialidad, necesitamos criterios o baremos de valoración. Los criterios de calificación han de ser bien concebidos, definidos de forma explícita y aplicados de forma constante. Los criterios bien especificados contribuyen a asegurar que todo el mundo entienda lo que se está pidiendo.

Los criterios bien articulados y públicos que se utilizan para valorar las respuestas de los alumnos son tan necesarios como útiles ya sea para utilizar los

[1]Muchos de los ejemplos utilizados en este libro provienen de programas de evaluación estatales, especialmente aquellos realizados en California. Debido a su trabajo pionero en el diseño de marcos curriculares que reflejan actuales teorías de aprendizaje y currículo, ciertos estados ya han probado prototipos prometedores para una evaluación alternativa que se pueden adaptar al aula.

Cuadro 5.1
Programa de evaluación de California 1990
Historia-ciencias sociales, 11avo año (grade 11)
Guía de calificación: Tarea de rendimiento en grupo

	Nivel I **Insuficiente**	**Nivel II** **Suficiente**	**Nivel III** **Satisfactorio**	**Nivel IV** **Notable**	**Nivel V** **Sobresaliente**
Aprendizaje en grupo y en colaboración 20	1-4: Dependencia exclusiva en un solo portavoz. Poca interacción. Conversaciones muy breves. Algunos alumnos muestran desinterés y distracción.	5-9: Dependencia fuerte en los portavoces. Sólo una o dos personas participan de forma activa. Interacción esporádica. La conversación no se centra completamente en el tema.	8-12: Alguna capacidad de interacción. Al menos la mitad de los alumnos se consultan o presentan ideas. Lectura cuidadosa de documentos y capacidad de escuchar con atención. Algunas muestras de que se sopesan diversas alternativas.	13-16: Los alumnos se muestran hábiles en la interacción. Al menos 3/4 de los alumnos participan de forma activa. Discusión animada sobre la tarea.	17-20: La mayoría de los alumnos participan con entusiasmo. Se comparte la responsabilidad de la tarea. Los alumnos respetan las opiniones de los otros e incluyen referencias a otras opiniones o alternativas en la presentación y la respuesta. Las preguntas y respuestas demuestran previsión y preparación.
Razonamiento crítico 30	1-6: Demuestra entender poco y tener un nivel de comprensión limitado del alcance del problema o asuntos que se han de tratar. Emplea sólo lo básico de la información provista. Confunde los hechos con las opiniones al desarrollar un punto de vista. Saca conclusiones después de una simple ojeada a una o dos partes de la información. No toma en cuenta las consecuencias.	7-12: Demuestra sólo una comprensión general del alcance del problema. Se centra en un único tema. Sólo emplea la información provista. Es posible que incluya opiniones en los hechos al desarrollar una postura. Saca la conclusión después de una revisión limitada de los datos sin preocuparse en gran medida por las consecuencias.	13-18: Demuestra una comprensión general del alcance del problema y de más de uno de los temas incluidos. Emplea los puntos principales de la información procedente de los documentos y conocimientos personales al desarrollar una postura. Basa la conclusión en la examinación de información y demuestra alguna consideración respecto a las consecuencias.	19-24: Demuestra una comprensión clara del alcance del problema y de al menos dos de los temas centrales incluidos. Utiliza los principales puntos de la información procedente de los documentos y conocimientos personales que son relevantes y comunes al desarrollar una postura. Basa la conclusión en el análisis de los datos más contundentes. Considera al menos una acción alternativa y sus posibles consecuencias.	25-30: Demuestra una comprensión clara y precisa del alcance del problema y de las ramificaciones de los temas incluidos. Emplea toda la información de los documentos y amplios conocimientos personales que son realmente relevantes, precisos y consistentes en el desarrollo de una postura. Basa la conclusión en un análisis cuidadoso de los datos hallados, en una exploración de alternativas razonables y en una evaluación de las consecuencias.

Cuadro 5.1 *(continuación)*

	Nivel I **Insuficiente**	**Nivel II** **Suficiente**	**Nivel III** **Satisfactorio**	**Nivel IV** **Notable**	**Nivel V** **Sobresaliente**
Comunicación de ideas 20	1-4: Su postura es vaga. La presentación es breve e incluye afirmaciones generales no relacionadas. El punto de vista global del problema no está claro. Las afirmaciones tienden a desviarse o a desentonar.	5-9: Presenta una postura general y sin definir. Has sólo una organización mínima en la presentación. Utiliza generalizaciones para respaldar su postura. Enfatiza o destaca sólo un tema. Considera sólo un aspecto del problema.	10-12: Adopta una postura definida pero general. Presenta un argumento medianamente organizado. Utiliza términos generales con pocas evidencias que pueden no ser totalmente precisas. Trata un número limitado de temas. Ve el problema con un alcance un tanto limitado.	13-16: Adopta una postura clara. Presenta un argumento organizado con tal vez errores mínimos en las evidencias. Trata los temas más importantes y demuestra alguna comprensión de las relaciones entre ellos. Considera la examinación de más de una idea o aspecto del problema.	17-20: Adopta una postura fuerte y bien definida. Presenta un argumento persuasivo bien organizado con evidencias puntuales. Trata todos los temas importantes y demuestra una profundidad de comprensión de las relaciones principales. Examina el problema desde varias posturas.
Conocimiento y utilización de la historia 30	1-6: Recita uno o dos hechos sin completa precisión. Trata los conceptos o temas sólo de forma breve y vaga. Apenas demuestra conocimientos previos de historia. Tiene una fuerte dependencia en la información dada.	7-12: Presenta sólo hechos básicos con poca precisión. Se refiere a la información para explicar al menos un tema o concepto en términos generales. Utiliza con limitada de conocimientos previos de historia sin precisión. Importante dependencia en la información dada.	13-18: Relaciona sólo los hechos importantes con los temas básicos con alguna precisión. Analiza la información para explicar por lo menos un tema o concepto con un apoyo sustancial. Utiliza ideas generales de sus conocimientos previos con algo de precisión.	19-24: Ofrece un análisis preciso de los documentos. Presenta hechos y los relaciona con temas principales pertinentes. Utiliza conocimientos previos de historia para analizar los temas implicados.	25-30: Ofrece un análisis preciso de la información y temas. Proporciona una variedad de hechos para explicar los temas y conceptos de mayor y menor importancia. Emplea extensivamente conocimientos previos para llegar a una comprensión profunda del problema y para relacionarlo con situaciones pasadas y del posible futuro.

resultados en el aula o para tomar decisiones a nivel centro o nacional. En todas estas situaciones de evaluación los criterios de calificación deben:

- Ayudar al maestro a definir un rendimiento sobresaliente y a planificar cómo ayudar a los alumnos a conseguirlo.
- Hacer saber a los alumnos lo que constituye un rendimiento sobresaliente y cómo evaluar sus propios trabajos.
- Comunicar a padres y otros cuáles son los objetivos y resultados.
- Ayudar a maestros y a otros calificadores a ser precisos, objetivos y constantes a la hora de calificar.
- Documentar los procedimientos utilizados al formular juicios importantes sobre alumnos.

Los criterios y la planificación docente

Los criterios de calificación aclaran los objetivos didácticos. Junto con la descripción de la tarea, los criterios definen los objetivos prioritarios en términos de contenidos que se ha de cubrir, conocimientos o destrezas que se ha de demostrar y contexto en el que van a surgir. Las especificaciones completas de evaluación alternativa pueden guiar la selección y la secuencia de las actividades didácticas relevantes.

Los criterios y los alumnos

Los criterios para las evaluaciones alternativas se hacen frecuentemente públicos para poder tratarse con los alumnos. Estas discusiones públicas ayudan a los alumnos a asimilar los criterios y "reglas" que necesitan para conseguir autonomía en sus estudios. Las evaluaciones alternativas y sus criterios pueden intercalarse dentro del mismo currículo de manera que sean claros para los alumnos y se perciban como una parte natural del proceso de aprendizaje. Tal evaluación es continua y cobra muchas formas—diarios, tutorías, clases de atención especial con maestros u otros alumnos, críticas de productos y exposiciones, y evaluaciones formales de trabajos individuales o de un cuerpo de trabajo. Los ejemplos de lo que constituye un buen trabajo alientan a los alumnos en el trabajo mismo y en las valoraciones de sus trabajos. El tratar públicamente los temas de calidad y criterios prepara a los alumnos durante el período formativo de enseñanza, no simplemente al final de una unidad o curso, cuando ya es demasiado tarde para llevar a cabo mejoras. Además, el tratar los criterios también ayuda a los alumnos a ver las perspectivas de sus maestros, de sus compañeros y, algunas veces, incluso de los expertos en el campo.

Los criterios y la participación de los padres

Los criterios articulados de forma clara también comunican a los padres y a otros aquello que los maestros y los centros pretenden lograr. Los criterios llevan a la práctica los objetivos de aprendizaje y las expectativas que se tiene de los niños. Cuando los padres saben antes de que los maestros califiquen a sus hijos lo que se espera de ellos, pueden ayudar en su aprendizaje. Por ejemplo, al proporcionar a los padres o educadores de centros preescolares una copia del "Perfil de Objetivos del Desarrollo en los Centros Preescolares -(Cuadro 5.2)" ("Profile of Developmental Outcomes for Kindergarten") se les permite trabajar en casa con sus hijos en actividades como son reconocer las primeras letras de las palabras o palabras fáciles de reconocer visualmente. El camino hacia la alfabetización está bien señalado; los maestros que comparten el recorrido con los padres verán que muchos de sus alumnos alcanzan su destino más rápidamente.

Los buenos criterios ayudan tanto a alumnos como a padres a compartir un segmento de la responsabilidad del aprendizaje. Hay menos probabilidad de que los padres y los niños que están familiarizados con los criterios que se utilizan para valorar el trabajo atribuyan el rendimiento insuficiente a factores externos como el no estar informados o a conflictos de personalidad entre maestros y alumnos.

Los criterios y la constancia

Cuando las directrices sobre lo que constituye un buen trabajo son vagas o no han sido establecidas, es difícil ser constante, justo y preciso a la hora de valorar las respuestas de los alumnos. En exámenes con respuestas a elegir, la precisión y la constancia en la calificación hacen referencia a si la nota de un determinado alumno se mantiene estable entre un examen y otro, en ausencia de enseñanza o desarrollo durante este período intermedio. Esta constancia se conoce mejor como fiabilidad. Para las evaluaciones alternativas, la fiabilidad no sólo incluye la idea de estabilidad de un determinado alumno a lo largo del tiempo, sino también la fiabilidad de las calificaciones del calificador de ese rendimiento. Para ser más específicos, una evaluación fiable que depende del juicio humano debe reunir los siguientes requisitos:

- Varios evaluadores que observasen una tarea específica deben llegar a la misma conclusión sobre un alumno.
- Cada evaluador calificaría el rendimiento del alumno en una tarea determinada de igual manera en exámenes consecutivos.
- El alumno llevaría a cabo la tarea de igual forma en distintas ocasiones.
- Si se pretende que la tarea represente o generalice alguna área superior, la muestra debe ser representativa de esa área.

Cuadro 5.2
Perfil de objetivos del desarrollo de las destrezas de capacidad de lectura y de cálculo en los centros preescolares

Joan C. Hillard, Superintendente, Spreckels Union School District, Spreckels, California
Elizabeth Jones, Catedrática, Pacific Oaks College, Pasadena, California
Jane Meade-Roberts, Directora y Propietaria, Power of Play Preschool, Salinas, California
San Vicente School, Soledad Union School District, Soledad, California (Jones y Meade-Roberts 1990)

Lenguaje oral	No habla en la escuela	Utiliza el lenguaje para satisfacer sus deseos y necesidades básicas	Utiliza mucho el lenguaje cuando juega y habla con sus compañeros	Describe claramente situaciones reales o imaginadas y utiliza un lenguaje descriptivo complejo	Habla con frases completas y utiliza un vocabulario bien desarrollado
Dibujo	Hace garabatos	Dibuja una cara	Añade brazos y piernas	Añade el cuerpo con brazos/piernas	Añade detalles (pelo, oreja, manos, etc.)
Expresión escrita	Hace garabatos y finge escribir	Utiliza letras o símbolos parecidos a letras para representar la escritura	Escribe de forma espontánea su propio nombre e incluye todas las letras	Copia palabras espontáneamente	Sabe deletrear palabras y utiliza la fonética
Lectura	Lee su propio nombre	Reconoce la letra inicial de su nombre cuando está escrita en otros sitios	Reconoce su propio nombre, otras letras y números	Reconoce y lee palabras familiares, entre ellas signos, etiquetas, palabras clave, listas de palabras creadas por el maestro y/o palabras en libros	Utiliza el conocimiento de los sonidos de las letras para intentar pronunciar palabras

Cuadro 5.2 ***(continuación)***

Actitudes hacia la alfabetización	Todavía no muestra interés por los libros o por escribir	Demuestra gran interés por los libros de dibujos	Demuestra interés por el lenguaje escrito (ej.: pregunta sobre signos, nombres, palabras en clase, etiquetas, palabras en libros o los sabe leer)	Practica escribir letras y números de forma espontánea	Demuestra interés por escribir correctamente
Resolución de problemas utilizando la clasificación	Manipula objetos seleccionados al azar	Ordena espontáneamente según semejanzas y diferencias	Reconoce o crea secuencias simples (AB) utilizando una variedad de materiales y/o símbolos	Reconoce o crea secuencias complejas (ej.: AABAAB) utilizando una variedad de materiales y/o símbolos	Puede clasificar utilizando más de un atributo a la vez (ej.: tamaño y color)
Resolución de problemas utilizando números	Cuenta al azar	Cuenta de memoria	Demuestra comprender la correspondencia de una cosa con otra (ej.: evalúa los objetos correctamente)	Es capaz de utilizar su conocimiento de contar para resolver problemas reales	Demuestra el sentido de permanencia de los números (ej.: comprende que el número de objetos permanece constante)

Cuadro 5.2 ***(continuación)***

Curiosidad	Observa en silencio	Hace preguntas con cautela	Formula preguntas constantemente	Formula preguntas cuando es debido	Utiliza recursos para buscar respuestas a preguntas (ej.: experimentando, arriesgándose, resolviendo problemas)
Creatividad	Espera a que le digan lo que tiene que hacer	Explora los materiales disponibles	Inventa una simple dramatización o proyectos con los materiales proporcionados	Pide o busca materiales no disponibles para lograr un proyecto/idea para un juego	Trabaja de forma hábil en tareas propias notablemente complejas, creativas e imaginativas
Destrezas sociales con sus compañeros	Normalmente observa a los otros jugando	Normalmente juega solo o juega paralelamente	Está adquiriendo destrezas de cooperación en el juego	Se siente seguro socialmente: juega bien con otros niños	Tiene destrezas bien desarrolladas de liderazgo y cooperación en el juego
Destrezas sociales con adultos/grupos	Acepta situaciones en lugar de buscar la ayuda de los adultos	Comunica con los adultos principalmente para solicitar ayuda	Habla de forma espontánea y libre con los adultos	Participa en actividades de grupo y en conversaciones	Es sensible a las necesidades de otros y sabe articularlas

Cuadro 5.2 *(continuación)*

Destrezas motoras generales	Corre	Salta	Salta con una sola pierna	Coge balones con brazos y pecho	Puede coger balones con sólo las manos
Destrezas motoras particulares	Hace garabatos con ceras/lápices	Sabe utilizar tijeras	Colorea dentro de las líneas y corta por las líneas	Dibuja/escribe líneas precisas	Siempre presenta trabajos limpios
Aprendizaje nuevo	Elige observar	Prefiere tareas familiares	Dispuesto a intentar tareas nuevas	Domina tareas nuevas rápidamente	Domina tareas nuevas independientemente
Conocimientos sociales	Conoce los colores	Conoce las formas	Conoce información personal	Conoce los nombre de las letras y números	Conoce los días de la semana y los meses
Tiempo de concentración	Cambia rápidamente	Se concentra en tareas seleccionadas por él/ella mismo/a	Se concentra en tareas seleccionadas por el maestro/a	Trabaja independientemente en tareas seleccionadas por él/ella y por el maestro/a	Puede seguir instrucciones complejas y concentrarse durante períodos largos

(De E. Jones y J. M. Roberts, *Profile of Developmental Outcomes for Kindergarten, Literacy and Numeracy Skills*, San Vicente School, Soledad CA)

Es evidente que estos cuatro requisitos para una calificación fiable exigen un mecanismo para crear el acuerdo entre los calificadores y para delimitar claramente las áreas de determinadas tareas de evaluación. Los criterios de evaluación deben responder a esta exigencia.

Los criterios y las consecuencias

Es siempre importante especificar los criterios y lo es aún más cuando las consecuencias de una evaluación son determinantes, por ejemplo, cuando las calificaciones pueden significar la repetición de un curso, o la graduación de un alumno, o la previsión de programas de atención especial. Unas directrices claras para evaluar el trabajo de un alumno aseguran consecuencias apropiadas para los alumnos y para el sistema educativo en conjunto. Además, cuando las evaluaciones alternativas se utilizan para estas decisiones determinantes, los procedimientos de calificación y los criterios deben poder defenderse ante un tribunal y deberán estar de conformidad con los procedimientos de dicho tribunal.

Especificar los criterios

Las distintas finalidades de los exámenes requieren distintos tipos de criterios de calificación. Muchos de los ejemplos de este libro fueron diseñados para evaluaciones a nivel estatal que conllevaban objetivos evaluadores trascendentales como la comparación de distintos centros, la identificación de centros que no funcionan al debido nivel y la evaluación de un centro determinado. Los criterios de un trabajo en grupo de la asignatura de historia (véase cuadro 5.1) del "Programa de Evaluación de California" (CAP—California Assesment Program) son un ejemplo de los complejos criterios utilizados en evaluaciones determinantes. Puesto que los criterios se utilizan en una evaluación final a nivel estatal, las directrices de calificación fueron desarrolladas para extraer la máxima información posible durante el tiempo limitado de la evaluación. Podemos observar que los criterios:

- Enumeran múltiples objetivos de aprendizaje.
- Dividen cada objetivo en niveles de rendimiento.
- Describen rasgos/características para cada nivel.
- Proporcionan una escala numérica para calificar el grado de alcance para cada nivel.
- Evalúan la calidad del rendimiento del alumno representado por los distintos niveles, utilizando descripciones como "insuficiente" o "sobresaliente".

Los criterios serán menos complejos cuando los objetivos de evaluación sean más centrados y las decisiones que se quieren tomar acerca de los alumnos sean

limitadas. Si se está utilizando diarios académicos de los alumnos para controlar y supervisar su progreso sobre cómo relacionar lo que aprenden en ciencias naturales con la vida real, los criterios de calificación podrían consistir en contar el número de frases espontáneas que relaciona el aprendizaje en el aula con la experiencia fuera del aula. El número de relaciones que se encuentren indicará si se están alcanzando los objetivos. La finalidad de la evaluación en este caso puede ser formativa—para mejorar la enseñanza y para identificar a aquellos alumnos que necesitan más ayuda o un tratamiento distinto.

Quizás la finalidad de la evaluación sea más tradicional—por ejemplo, se quiere evaluar el progreso del alumno con referencia a los objetivos de resolución de problemas matemáticos. Los criterios de calificación podrían imitar la rúbrica generalizada diseñada por el CAP para problemas matemáticos que requieren una redacción desarrollada (véase cuadro 5.3). Los criterios proporcionan descripciones de cada nivel del rendimiento en términos de lo que los alumnos sean capaces de hacer, atribuyen valores para estos niveles, después aplican parámetros en determinados puntos para distinguir los distintos niveles. Los alumnos que reciben una calificación entre 1-2 son aquellos que han tenido una respuesta "inadecuada"; los alumnos que reciben entre 3-4 se consideran "aptos"; y a los alumnos que reciben entre 5-6 se les considera "competentes".

Aunque la calificación es un tema complejo y la calificación de cualquier evaluación alternativa puede o no utilizarse para decidir las notas finales, es posible encontrar o establecer criterios relacionados con notas que corresponden a determinadas letras. Gracias a un subsidio de la National Science Foundation, los investigadores han formulado un conjunto de criterios que corresponden a letras para evaluar los conocimientos de alumnos sobre los procedimientos científicos en un experimento científico práctico (Baxter et al. 1992). Los investigadores establecieron cuáles eran los métodos que los alumnos podrían utilizar para resolver el problema planteado en el experimento y juzgaron cuál de ellos produciría las soluciones más lógicas y eficaces. Luego establecieron criterios con referencias alfabéticas para reflejar sus valoraciones de las soluciones. Un resumen de sus criterios utilizando letras se describe en el cuadro 5.4.

Sin tomar en cuenta la finalidad de la evaluación, los criterios que se describen tienen cuatro características en común. Cada uno tiene:

- Uno o más rasgos o **dimensiones** que constituyen la base de la valoración de la respuesta del alumno.
- **Definiciones y ejemplos** para aclarar lo que significa cada rasgo o dimensión.
- Un **baremo** de valores (o un sistema numérico) que se sigue para calificar cada dimensión.
- **Estándares** idóneos para determinados niveles de rendimiento acompañados por modelos o ejemplos de cada nivel.

Cuadro 5.3
Rúbrica generalizada del CAP
(Departamento de Educación del Estado de California 1989)

Competencia Demostrada

Respuesta ejemplar...Calificación = 6
Da una respuesta completa con una explicación clara, coherente, sin ambiguedades y elegante; incluye un diagrama claro y simplificado; se comunica eficazmente con su público; demuestra entender las ideas y los procesos del problema matemático de solución abierta; identifica todos los elementos importantes del problema; posiblemente incluya ejemplos y contraejemplos; presenta fuertes argumentos respaldatorios.

Respuesta competente...Calificación = 5
Da una respuesta bastante completa con explicaciones razonablemente claras; posiblemente incluya un diagrama apropiado; se comunica eficazmente con su público; demuestra entender las ideas y procesos matemáticos del problema; identifica los elementos más importantes de los problemas; presenta sólidos argumentos respaldatorios.

Respuesta Satisfactoria

Pequeños fallos aunque satisfactoria...Calificación = 4
Termina el problema de forma satisfactoria, aunque la explicación puede ser confusa; los argumentos pueden aparecer incompletos; el diagrama puede ser inapropiado o no estar muy claro; entiende la ideas matemáticas subyacentes; utiliza ideas matemáticas con eficacia.

Fallos serios aunque casi satisfactoria...Calificación = 3
Empieza correctamente el problema pero puede terminarlo mal u omitir partes importantes del problema; puede fallar en la comprensión completa de las ideas y de los procesos matemáticos; posiblemente cometa importantes errores de computación; posiblemente utilice mal o no utilice términos matemáticos; la respuesta puede reflejar una estrategia inapropiada para la resolución del problema.

Respuesta insuficiente

Empieza pero no termina el problema...Calificación = 2
La explicación no se entiende; el diagrama podría no estar muy claro; demuestra no entender el problema; posiblemente cometa importantes errores de computación.

Incapaz de abordar el problema eficazmente...Calificación = 1
Lo que escribe no refleja el problema; los dibujos representan mal el problema; copia partes del problema sin tratar de resolverlo; no indica cuál es la información apropiada para el problema.

No lo intenta...Calificación = 0

Cuadro 5.4
Relación criterios-notas

Nota	Criterios para determinar las notas
A	El alumno selecciona el método. El alumno empapa las toallas. El alumno comprueba el resultado para contestar la pregunta. El resultado lógicamente corresponde al método utilizado para empapar la toalla. Las medidas se toman con precisión/cuidado. Las conclusiones son correctas.
B	Cumple todos los requisitos del "A" excepto que las medidas no se toman con cuidado.
C	Cumple todos los requisitos del "A" aunque comete algunas equivocaciones. Debe intentar controlar el proceso de empapar poniendo la misma cantidad de agua en cada toalla. Las toallas no están empapadas (la dimensión clave para decidir otorgar una "C" u otra inferior).
D	El alumno no empapa las toallas ni la toalla de control. El resultado es lógicamente falto de uniformidad debido al método utilizado para empapar las toallas.
F	El alumno no llevó a cabo la investigación O el equipo se utilizó sin ningún propósito O las toallas no estaban mojadas O las conclusiones se basaron en el cambio que sufrieron las toallas.

* Criterios abreviados de Baxter et al. (1992, p. 5).

Consideraciones al selecionar las dimensiones

Las dimensiones que se utilicen para evaluar el rendimiento de un alumno en una determinada área deben reflejar las cualidades esenciales de un buen rendimiento en esa área. ¿Dónde se encuentran estas cualidades esenciales? Las cualidades o dimensiones pueden marcarlas expertos fuera del campo de la educación, colegas del mismo departamento, maestros de distintos cursos, comités de programas de estudio a nivel distrito, artículos de investigación, o comités de estándares de distintas asignaturas de la localidad. Si lo que se quiere es establecer criterios para

la propia aula, éstos se deben centrar en aquellos aspectos del rendimiento del alumno que reflejan los objetivos didácticos de mayor prioridad y que representan aspectos del rendimiento que se pueden enseñar y observar.

Al formularse el siguiente tipo de preguntas, se puede descubrir dimensiones para crear criterios de evaluación:

- ¿Cuáles son los atributos de una buena redacción, de un buen razonamiento científico, de un buen proceso colaborativo, de una presentación oral acertada? ¿En general, qué cualidades o rasgos me indicarán si los alumnos han respondido de forma óptima a la tarea de evaluación?
- ¿Cómo se relaciona la realización de esta tarea con los objetivos establecidos para los alumnos? ¿Qué harán para demostrar que se está alcanzando o logrando alguno de estos objetivos?
- ¿Qué se espera ver si se realiza esta tarea de manera óptima, aceptable, insuficiente?
- ¿Se dispone de ejemplos o modelos de trabajos de alumnos, de la misma clase o de otras fuentes, que ejemplifiquen algunos de los criterios que se podrían emplear cuando se evalúe esa tarea?
- ¿Qué criterios existen para esta tarea u otras parecidas en las directrices de programas de estudio estatales, en el programa de evaluación del estado, en las guías de programas de estudio del distrito o en el programa de evaluación del centro?
- ¿Qué dimensiones se pueden adaptar del trabajo que realizaron los consejos nacionales de programas de estudio, u otros maestros?

Además de describir la valoraciones que hacemos sobre el rendimiento, hay que describir las dimensiones que se van utilizar para los criterios de tal forma que todos aquellos que las van a utilizar las entiendan de igual manera. Quizás se desea evaluar un proyecto de arte interdisciplinario. Por ejemplo, diseñado donde se refleje una interpretación-n desde el punto de vista de las ciencias sociales-n de la relación entre los pueblos indígenas americanos y su medio ambiente. Los criterios para calificar o evaluar niveles de rendimiento deben ser claros tanto para los alumnos como para los padres. También deben ser claros para otros maestros que dependen de estas valoraciones sobre el dominio del contenido, ya sea otros maestros que dicten el mismo curso o que lo dictarán en un futuro.

Hay varias maneras de llegar a descripciones claras sobre las dimensiones del rendimiento:

1. Se puede escribir definiciones referentes a las conductas o elementos que se verán a la hora de evaluar a los alumnos. Por ejemplo, en lugar de decir "Un rendimiento aceptable significa que los alumnos demuestran entender la idea de vivir en armonía con la tierra" se podría decir "Un rendimiento aceptable significa que los dibujos del alumno muestran un medio ambiente que apenas ha cambiado desde sus orígenes. Pocos árboles han sido talados; la pradera se encuentra intacta con la excepción

de pequeñas parcelas que han sido cultivadas; no existen grandes vertederos, etcétera".

2. Se puede proporcionar modelos o ejemplos para cada dimensión. Esto se suele hacer en evaluaciones directas de la expresión escrita. Se da a los maestros copias de redacciones de alumnos que ejemplifican cada punto en la distribución de la calificación. Las redacciones ilustran dimensiones como: "la redacción está bien estructurada; comienza y termina bien". Con estos modelos, los maestros y otros pueden llegar a articular definiciones precisas para cada dimensión.
3. Si la evaluación es informal, se puede definir las dimensiones con un conjunto de preguntas. Por ejemplo, al evaluar diarios para determinar el tipo de ayuda que necesitan los alumnos para adquirir la soltura en la expresión escrita, los criterios para decidir en qué debemos trabajar a continuación pueden incluir las siguientes preguntas: ¿Cuáles alumnos están utilizando estrategias previas a la redacción como son agrupar ideas, dibujar, hacer listas, o apuntar ideas sueltas? ¿Cuáles alumnos están llevando un diario de ideas para futuras redacciones? ¿Cuáles alumnos tienen problemas ortográficos que bloquean el flujo de ideas?

Las definiciones de calificaciones que no sufren ambigüedades normalmente consisten en una descripción de las dimensiones que se han de evaluar, junto con modelos de trabajos de alumnos que muestran respuestas aceptables. Estos modelos o ejemplos de trabajos son cruciales cuando se desea llegar a un consenso sobre el significado de los criterios cuando se utilizan en la formación de calificadores para evaluaciones formales. Los modelos también proporcionan a los alumnos ejemplos concretos de lo que es un trabajo aceptable o excelente. El cuadro 5.5 muestra una de las muchas dimensiones de una rúbrica de calificación desarrollada por el CRESST para evaluar la redacciones de los alumnos de enseñanza secundaria para determinar el grado de comprensión de la asignatura de historia. Véase que las dimensiones y las calificaciones son totalmente prácticas: se define palabras clave como "concepto" y se proporciona ejemplos de elementos básicos, como afirmaciones de opinión.

En la mayoría de los casos, las dimensiones de rendimiento, en particular aquéllas para la evaluación en el aula, reflejarán nuestras opiniones de lo que constituye la excelencia o dominio y se verán moderadas por nuestras expectativas acerca de los alumnos de distintos cursos y por nuestros objetivos didácticos en distintas épocas del curso escolar. Puesto que los criterios ayudan a los alumnos a concentrarse en lo que es importante desde el punto de vista académico, se puede utilizar distintos criterios en distintos momentos del curso escolar. Por ejemplo, aunque se considera que la organización y los mecanismos son aspectos importantes en la expresión de conocimientos con relación a las disciplinas de ciencias o historia, quizá al comienzo del curso escolar se pueda hacer hincapié en la soltura. Por consiguiente, los criterios al comienzo del semestre pondrán énfasis en el número de ideas presentadas, el número de ejemplos o definiciones para cada idea, etcétera. Cuando los alumnos adquieran mayor soltura y sean capaces de respaldar sus opiniones, se puede ampliar los criterios para incluir la organización

Cuadro 5.5
Explicación del área de contenido CRESST
Directrices de calificación de redacción
(Baker, Aschbacher, Niemi y Sato 1992)

Rúbrica de Calificación CRESST:

Impresión general—calidad de contenido
Número de principios o conceptos
Conocimientos previos: hechos y acontecimientos
Argumentación
Ideas equivocadas
Detalles textuales

Ejemplos de directrices para la Escala *Número de Principios o Conceptos:*

Número de principios/conceptos

Esta es una valoración del número de distintos conceptos o principios de ciencias sociales que utiliza el alumno demostrando que los entiende.

Un *concepto* es una noción general abstracta, como lo es la "inflación". No se refiere a objetos o a acontecimientos determinados (como un período de inflación particular), sino que representa rasgos comunes de una categoría de acontecimientos u objetos. El "imperialismo", por ejemplo, no se refiere a hechos o acontecimientos específicos sino que es un nombre que define una clase de conductas y creencias. De igual manera, la "industrialización" identifica una clase de actividades y acontecimientos que comparten propiedades comunes. Hay que asegurarse de que el alumno esté utilizando un término de forma conceptual y no simplemente como etiqueta.

Un *principio* es una regla o creencia que se utiliza para justificar una acción o juicio, como en el enunciado "La esclavitud es inmoral", donde la "moralidad" sirve como un principio que justifica.

Debe quedar claro que el alumno entienda el concepto y que sea su intención discutirlo. El concepto no debería simplemente mencionarse dentro de una cita del texto sin ninguna indicación de que el alumno lo entiende. Para ganar puntos, no hay que nombrar de forma explícita el concepto o principio como en la frase "La constitucionalidad era un principio importante que influyó en el debate sobre la esclavitud", sino que hay que enunciar la idea claramente, por ejemplo, "Un problema era determinar lo que decía la constitución acerca de la esclavitud".

Directrices de calificación:
0—no responde
1—ningún concepto/principio
2—un concepto/principio
3—dos conceptos/principios
4—tres conceptos/principios
5—cuatro o más conceptos/principios

Ejemplo: "Un factor importante que impidió que fuéramos a la guerra fue nuestra economía. No se sabía lo que ocurriría con nuestra economía sin la seguridad que nos daba Gran Bretaña. Gran Bretaña podía defender nuestro comercio y costas. Además, con Gran Bretaña teníamos la ventaja de la exportación. Parecía que nuestra economía sólo iba a sufrir si prescindíamos de la ayuda de Gran Bretaña".

y los mecanismos de la escritura. Tomando un ejemplo del patinaje artístico, es posible que creamos tanto en los criterios olímpicos de "mérito técnico" como en los de "expresión artística", pero en distintas fases de la enseñanza habrá que cambiar el énfasis de uno a otro.

Dimensiones para tareas complejas

Como ya establecimos en el capítulo 4, es perfectamente posible crear una evaluación compleja con múltiples finalidades distintas. Al tener muchas finalidades, se requiere de muchos criterios, un conjunto para cada finalidad. No se puede evitar los criterios multidimensionales cuando se está llevando a cabo una evaluación interdisciplinaria o cuando estamos evaluando objetivos de aprendizaje complejos. Se puede formular criterios independientes para cada una de las finalidades, o bien, formular un conjunto multidimensional de criterios. La evaluación del estado de Connecticut para las ciencias utiliza un enfoque bipartita para evaluar la misma tarea al proporcionar criterios que evalúan el proceso en grupo y el logro individual (véase cuadros 5.6 y 5.7). Otra perspectiva sobre el rendimiento del alumno la proporcionan las subdestrezas dentro de las evaluaciones individuales y de grupo. Al valorar las destrezas de proceso en grupo, estamos interesados en el proceso científico, la comunicación y la colaboración en grupo. Hay criterios independientes para cada una de estas destrezas. Las dimensiones múltiples en la escala individual incluyen finalidades de contenido y de comunicación.

Las dimensiones para cada escala requieren de mucha inferencia. Tanto maestros como alumnos necesitarían más descripciones de dimensiones como "sacar conclusiones razonables" o "colaborar con eficacia" para poder utilizar los baremos. De hecho, estos baremos se utilizan en el aula sólo después de que los maestros hayan recibido una formación interna para tratar el significado de las dimensiones, de los ejemplos y para poder practicar utilizando los criterios. Cuando se ventila en el aula estos temas y se presentan los ejemplos, los alumnos y maestros llegan a comprender mutuamente las dimensiones del baremo individual.

Un ejemplo menos complejo de los criterios multidimensionales se encuentra en el cuadro 5.1. Los criterios evalúan cuatro objetivos del rendimiento en grupo: la colaboración, el razonamiento crítico, la comunicación y los conocimientos de historia. Los criterios incluyen subcriterios para decidir en cuál de los cinco niveles de rendimiento debemos colocar a los alumnos según cada objetivo. El conjunto completo de criterios de trabajo en grupo puede verse como un compendio de cuatro conjuntos de criterios: para la colaboración, para el razonamiento crítico, para la comunicación y otro para los conocimientos de historia.

63

Cuadro 5.6
Parte II: Formulario de calificación de objetivos—grupo

Nombre alumnos:
1. _____
2. _____
3. _____
4. _____
5. _____

Título de la tarea: _____ Tarea: _____

Nombre maestro: _____ Fecha: _____

El grupo debería ser capaz de...	Dónde encontrar evidencias: Informe de grupo (# de pág.)	Presentación oral	Observación del maestro	Otro (señalar)	S	B	N.M.	*
1. Identificar y aplicar propiedades físicas y/o químicas con el fin de la identificación.					○	○	○	
2. Formular predicciones basadas en conocimientos previos.					○	○	○	
3. Identificar la información y los pasos necesarios para resolver un problema.					○	○	○	
4. Comprobar sus predicciones.					○	○	○	
5. Recopilar datos pertinentes a un problema.					○	○	○	
6. Deducir basándose en los datos pertinentes.					○	○	○	
7. Sacar conclusiones razonables y defenderlas racionalmente.					○	○	○	
8. Comunicar las estrategias y resultados de un estudio mediante la expresión escrita.					○	○	○	
9. Comunicar oralmente las estrategias y resultados de un estudio.					○	○	○	
10. Colaborar con eficacia.					○	○	○	
* Compruébese si el trabajo del alumno es un ejemplo bueno y claro de la calificación dada.					○	○	○	

(Connecticut Department of Education 1990) S = Sobresaliente B = Bien N.M. = Necesita Mejorar

Cuadro 5.7
Parte II: Formulario de calificación de objetivos—individual

Título de la tarea: _____ Tarea _____ Nombre alumno _____

Nombre maestro: _____ Fecha _____

El grupo debería ser capaz de...	Dónde encontrar evidencias				S	B	N M	*
1. Identificar y aplicar propiedades físicas y/o químicas con el fin de la identificación.					○	○	○	
2. Identificar la información y los pasos necesarios para resolver un problema.					○	○	○	
3. Comunicar las estrategias de un estudio mediante la expresión escrita.					○	○	○	
					○	○	○	
					○	○	○	
					○	○	○	
					○	○	○	
					○	○	○	
					○	○	○	
* Compruébese si el trabajo del alumno es un ejemplo bueno y claro de la calificación dada.								

(Connecticut Department of Education 1990) S = Sobresaliente B = Bien N M = Necesita Mejorar

Utilizar los baremos de calificación

Todos los ejemplos de criterios de calificación que se incluyen en este capítulo contienen algún tipo de baremo, bien numérico, cualitativo, o ambos. Los criterios en el cuadro 5.1, el trabajo en grupo de historia y en el cuadro 5.3, el problema de matemáticas, contienen tanto baremos de calificación numéricos como cualitativos. El cuadro 5.4, que comprende los criterios científicos prácticos, y los cuadros 5.6 y 5.7, del experimento de ciencias en grupo e individual, sólo contienen calificaciones cualitativas, como por ejemplo notas en forma de letras o valoraciones tales como "sobresaliente" o "necesita mejorar".

¿Por qué se usan los baremos? ¿Cómo se sabe si se debe emplear calificaciones numéricas o cualitativas? ¿Y qué tal si empleamos una lista de control en lugar de un baremo de calificación? Ya se puntúe la presencia o ausencia de un rendimiento, como en una lista de control, o se utilice números o evaluaciones cualitativas, todo dependerá del objetivo de la evaluación. Hay tres tipos principales de baremos: listas de control, calificación numérica y calificación cualitativa (ya sea descriptiva o evaluativa). Si el objetivo es **describir** lo que pueden hacer los alumnos, por ejemplo para reuniones con los padres o para comparar el rendimiento del alumno con ciertos estándares de desarrollo, se puede utilizar el baremo de calificación más simple de todos, la lista de control. Si hace falta más información aparte del simple hecho de que si un alumno se ocupa de aspectos específicos de una tarea, se necesitará un baremo de calificación totalmente desarrollado. Si se quiere averiguar **hasta qué punto** fueron observadas las dimensiones o la **calidad** del rendimiento, hacen falta baremos más elaborados. Los baremos de calificación, más allá del formato del "sí o no" de una lista de control, reflejan distintos aspectos del rendimiento del alumno y no simplemente los logros alcanzados en una determinada actividad.

Listas de control

Una lista de control es una lista de dimensiones, características, o conductas que se valoran simplemente con un "sí o no". Un control indica que bien la característica o la conducta estaba presente o ausente. Las listas de control contienen muchas veces más dimensiones para calificar de las que contienen los baremos de calificación, pero estas dimensiones son frecuentemente limitadas y concretas.

Las listas de control pueden ser útiles en la valoración de procesos, un importante objetivo para los maestros que se preocupan por el cómo además del por qué del aprendizaje. Una lista de control de procesos para un experimento práctico podría imitar el cuadro 5.8, que pide al calificador que anote la presencia de determinadas conductas.

Los maestros de educación primaria encuentran útiles las listas de control porque con frecuencia se emplean para saber cómo evolucionan los alumnos según alguna teoría de adquisición de destrezas. Por ejemplo, la actual teoría de la adquisición del lenguaje sugiere que el siguiente conjunto de destrezas apoya la

habilidad de leer de un niño:

- Habilidad de dibujar o representar una idea
- Habilidad de reconocer la relación entre sonidos y letras
- Habilidad de reconocer que las palabras representan algo
- Conocimiento de la orientación de la página de izquierda a derecha y de arriba a abajo
- Habilidad de recordar y repetir sus cuentos preferidos

Cuadro 5.8
Lista de control de procesos

Procedimiento	Conducta observada	Comentarios
Eligió un método		
Utilizó el material adecuado		
Las medidas fueron correctas		
Pidió ayuda a los compañeros cuando la necesitó		
Anotó observaciones		
Limpió al acabar el experimento		

El maestro puede documentar la adquisición de estas destrezas de agilidad mental con una lista de control. No es necesario juzgar lo bien que estas conductas se demuestran, sino simplemente señalar que existen. El cuadro 5.2 muestra un perfil basado en el desarrollo de niños de educación preescolar, creado por maestros del Soledad Union School District en California, con alguna colaboración del Pacific Oaks College en Pasadena, California. Este es un ejemplo de un perfil basado en teoría. El proceso del desarrollo del perfil fue diseñado para ayudar a los maestros a comprender mejor el constructivismo, la teoría del desarrollo del aprendizaje en la que se basa este perfil. Las conductas identificadas en el cuadro 5.2 siguen una secuencia de izquierda a derecha siguiendo el orden de adquisición de las conductas que predijeron los maestros del centro enseñanza preescolar. Este documento fue diseñado para que se vuelva a analizar cada año conforme los maestros vayan observando las conductas de los niños desde el punto de vista del desarrollo.

Baremos numéricos

Un baremo numérico utiliza números o asigna puntos en un espectro de niveles de rendimiento. La extensión del espectro o el número de puntos en el baremo puede variar, tres puntos, cuatro puntos, cinco puntos, siete puntos—cualquier número es posible. ¿Cuántas divisiones o puntos debiera incluir un buen baremo? Aunque no hay una única respuesta a esta pregunta, la experiencia nos dice que tomemos en cuenta estos temas.

El número de puntos o divisiones en un baremo puede y debe variar de acuerdo con las decisiones que se tomen con respecto a los alumnos y si el baremo se va a utilizar en el aula o en una sesión formal de calificación con varios calificadores involucrados en la evaluación del rendimiento. Generalmente, entre mayor es el baremo, más difícil será distinguir claramente entre los puntos. Consideremos lo rápido que resulta clasificar de redacciones en pilas según reciban cero puntos, un punto o dos puntos; ésta es esencialmente una decisión entre bajo, mediano y alto. ¿Por qué habría que utilizar un baremo de diez puntos si realmente sólo queremos distinguir entre dos o tres grupos de alumnos, como por ejemplo entre aquellos que necesitan ayuda especial para escribir una redacción bien estructurada y aquellos que no?

Un baremo con pocos puntos también tiene sus desventajas. Un mayor número de puntos nos ayuda a identificar pequeñas diferencias entre distintos alumnos y puede proporcionarnos más información diagnóstica que un baremo menor. Por ejemplo, puede ser necesario usar un baremo más detallado si lo que se desea es utilizar un único baremo para todos los alumnos de K–12 a la vez que se quiere diferenciar los alumnos de un sólo curso. Además, si el baremo se va a utilizar para fines de evaluación formales donde varios lectores van a calificar cada rendimiento, cualquier estadística que haya que calcular, como es el nivel de acuerdo entre los calificadores, se verá afectada por el tamaño del baremo. La utilización de un baremo más pequeño dará lugar a un alto porcentaje de acuerdo pero será más difícil alcanzar una mayor correlación entre las calificaciones de los calificadores (dos formas distintas de calcular la fiabilidad entre calificadores).

Se tarda más en llegar a un consenso sobre cómo asignar los puntos cuando se tiene que tener en cuenta un mayor número de ellos. Con un baremo de cinco a seis puntos, los calificadores con frecuencia recurren a experiencias anteriores y asignan los puntos más bajos a aquellos rendimientos irrelevantes o verdaderamente terribles, los más altos a aquellos que son brillantes, reservan los puntos intermedios para aquellos rendimientos que son "aptos", "aceptables", o modelo, para luego calificar a aquellos rendimientos que no se ajustan a las tres calificaciones base utilizando los valores de la escala que quedan. Una escala de once o diecisiete puntos hace que sea más difícil para los calificadores basar sus juicios en experiencias anteriores. Sin embargo, se encuentran con frecuencia baremos de múltiples de cinco, como por ejemplo baremos de diez, quince o veinte puntos, que permiten a los calificadores agrupar los puntos de cinco en cinco. Las distinciones iniciales de calificaciones se hacen entonces entre un cinco y un diez, en lugar de entre un cuatro y un siete, con los ejemplos de rendimiento que no se ajustan claramente a los incrementos que reciben los puntos intermedios.

Otra consideración relacionada con el tamaño del baremo hace referencia a los criterios multidimensionales. Si se califica el mismo rendimiento de acuerdo con varios criterios, donde cada uno evalúa un objetivo distinto, quizás se quiera utilizar el mismo número de puntos para cada objetivo. Esto no sólo logra que sea posible que se junten o se puedan comparar los resultados de varios baremos, sino que facilita la tarea de calificación. Por ejemplo, el uso de un baremo de cuatro puntos para valorar la coherencia y de otro de cinco puntos para valorar los argumentos respaldatorios puede hacer más lento el proceso de calificación sea más pues forzará a los calificadores a cambiar mentalmente a un baremo de calificación distinto. Los alumnos al intentar entender sus relativos puntos fuertes y débiles, también pueden encontrar dificultades al comparar distintos baremos. Sin embargo, si se quiere que determinados objetivos tengan más peso que otros en el cómputo total, se puede emplear baremos de distinta extensión para reflejar el valor o peso relativo. Un buen ejemplo de esta estrategia se muestra en el cuadro 5.1, la tarea de trabajo en grupo de historia. La guía de calificación utiliza dos baremos distintos; a un objetivo se le asigna veinte puntos y al otro treinta.

Baremos cualitativos

Un baremo cualitativo utiliza adjetivos en lugar de números para describir el rendimiento del alumno. Estos baremos son generalmente de dos tipos, descriptivos y evaluativos. Los descriptivos catalogan el rendimiento del alumno pero no necesariamente hacen explícitos los estándares subyacentes de las valoraciones hechas; utilizan términos bastante neutrales para describir el rendimiento. Entre los descriptores típicos tenemos las valoraciones sobre la realización de la tarea, la comprensión de la tarea o la presencia de determinados elementos en el rendimiento. El cuadro 5.9 muestra tres baremos descriptivos que no evalúan el valor del rendimiento del alumno.

Cuadro 5.9
Escalas descriptivas

Ninguna evidencia...Evidencia mínima...Evidencia parcial...Evidencia total

Tarea no realizada...Realización parcial...Realizada...Supera lo esperado

Irrelevante a la tarea...Intenta abordar la tarea...Atención mínima en la tarea...
Se entrenta a la tarea pero no la lleva a cabo...Llevada a cabo en su totalidad y
concentrado en la tarea y en el receptor.

Los baremos evaluadores incorporan juicios de valor basados en los criterios subyacentes de lo que se considera excelente. Los baremos evaluadores más

comunes son los de letras (véase cuadro 5.4). Los baremos que utilizan descriptores referentes a un posible rendimiento "sobresaliente" (cuadros 5.1, 5.6 y 5.7) o que evalúan la competencia (cuadro 5.3) son evaluadores por naturaleza. Los baremos evaluadores requieren de un mayor grado de inferencia que las baremos descriptivos para poder interpretarlos. Estas inferencias se hacen teniendo en cuenta los criterios de calificación. Los mismos criterios llevan consigo nociones de rendimiento sobresaliente, competencia o resultados aceptables.

Baremos numéricos-cualitativos

Los baremos numéricos son normalmente más fáciles de recordar, de calcular y de sacar la media, pero son difíciles de interpretar cuando no tienen buenos descriptores. Al fin y al cabo, sacar un "4" en un baremo de seis puntos puede significar niveles o cualidades de logro diferentes para distintas personas. Los buenos criterios incluyen normalmente tanto valores numéricos como descriptivos. Por ejemplo, el cuadro 5.3 muestra un borrador de un baremo utilizado por el California Assessment Program para calificar problemas matemáticos con solución abierta. Este baremo, como vemos, es tanto numérico como descriptivo. El rendimiento se evalúa de forma numérica, pero cada calificación numérica se une a una valoración que va de "insuficiente" a "competente".

Ya sean los valores del baremo numéricos, descriptivos, o ambos, es importante asegurar que los baremos ayuden a los padres, alumnos, maestros, personal administrativo y coordinadores educativos a comprender de igual manera el significado del rendimiento. Esta concordancia en la comprensión ayuda a garantizar la fiabilidad y la objetividad en las evaluaciones.

La relación con los estándares académicos

Casi todos los criterios, incluso las listas de control descriptivas, se relacionan de alguna manera con los estándares académicos—las expectativas del rendimiento del alumno. Las notas o calificaciones cualitativas reflejan el juicio del maestro, o en el caso de los criterios prácticos de ciencias que aparece en cuadro 5.4, reflejan el consenso del equipo calificador. Los criterios subyacentes de baremos distintos pueden reflejar bien métodos de evaluar la calidad referente a criterios o referentes a la norma. Los criterios para la asignatura de matemáticas (cuadro 5.3) con descriptores como "respuesta insuficiente", "respuesta satisfactoria" y "competencia demostrada" reflejan un nivel absoluto o un énfasis en el dominio a la hora de establecer los estándares deseados. Las descriptores indican claramente niveles de rendimiento buenos o deseados, "de satisfactorio en adelante", frente a niveles más pobres, "insuficiente". Los niveles tienen como referencia los estándares basados en la disciplina, los conceptos de los maestros de matemáticas sobre lo que constituyen estrategias adecuadas para la resolución de problemas.

Otro ejemplo lo encontramos en el baremo de seis puntos utilizado para evaluar la expresión escrita en Illinois, que emplea un baremo absoluto y está diseñado para poder utilizarse en distintos cursos escolares. Una calificación de seis representa un nivel muy alto en la expresión escrita y es de esperar que tan sólo unos pocos alumnos de educación primaria, si es que llega a haber alguno, superen la calificación de "3". Este tipo de baremo es especialmente útil para medir la evolución conforme pasan los años. La limitación de un baremo absoluto para la evaluación de varios cursos y edades se debe a que los alumnos de primaria suelen sacar las notas más bajas del baremo; hay poca variabilidad en sus notas por lo que es imposible, a partir de ellas, averiguar mucho sobre ellos de forma individual. Todos "se parecen".

Otros baremos evaluadores reflejan métodos que utilizan la norma como referente para el establecimiento de los estándares. Cuando se asignan calificaciones o puntos comparando el estatus relativo de los alumnos, como por ejemplo "la redacción de María estuvo por encima de la media de la clase", "el vídeo de Gary estuvo entre los mejores de la clase", los estándares utilizan la norma como referente. Las listas de control o baremos del desarrollo demuestran otro uso frecuente de los baremos que toman la norma como referente para la evaluación alternativa. La secuencia de las conductas en estos baremos depende de lo que los educadores y otros han ido observando a lo largo del tiempo como rendimiento típico en determinadas edades. Por ejemplo, los niños que alcanzan la "media" en lectura demuestran conductas típicas de su edad o de su curso. "Por debajo de la media" o "evoluciona con lentitud" se refieren al rendimiento típico de los niños de edad inferior al grupo de aquellos que se está evaluando.

Los estándares se pueden basar en la información que parte tanto de los criterios establecidos como de la norma para una misma evaluación. Se empieza con un baremo que utiliza a los criterios como referente, un baremo que describe el rendimiento con relación a un conjunto claramente definido de conductas. Luego se recopila, o se obtiene por otros medios, los datos sobre cómo realizaron la misma prueba los alumnos de una muestra representativa a nivel nacional, de estado o local. A continuación se podrían formular frases como "María escribió una redacción bien organizada y se le dio un "4" en la estructura; su rendimiento fue descrito como superior al 75 por ciento de los alumnos del estado". O a un nivel más informal, en el aula, siempre se puede describir el nivel de rendimiento de un determinado alumno comparándole con el resto de la clase: "La nota de María la coloca entre las mejores de la clase".

Algunos baremos pueden parecerse a los baremos de referente absoluto o de criterios establecidos aunque en realidad pueden incorporar tanto información referente a la norma como a los criterios. Un baremo relacionado con la edad o con el curso escolar define el rendimiento del alumno en términos de los parámetros o expectativas de un curso determinado. Los parámetros para la resolución de problemas matemáticos en 5º año ("5th grade") serán diferentes a los que se establezcan para el 7º año ("7th grade"). Lo que se considera excelente en la estructura de una redacción en el 8º año ("8th grade") no lo será en un 11avo año ("11th grade"). Aunque parezcan relacionados con criterios establecidos, los baremos ligados a cierta edad o al currículo de un determinado curso escolar se

pueden interpretar subyacentemente como baremos que utilizan la norma como referente. Las propias dimensiones se establecieron a partir de lo que los alumnos eran capaces de hacer en determinados cursos y no a partir de estándares absolutos de rendimiento de todas las edades y cursos. Por razones prácticas se considera que estos baremos por curso utilizan como referente criterios prefijados porque su primer objetivo es decidir lo que los alumnos son capaces de hacer con respecto a un contenido y destrezas determinados en lugar de compararlos mutuamente.

¿Cómo se puede conseguir lo mejor de ambos? Determinando los estándares apropiados prestando atención a los objetivos de la evaluación. Para la evaluación en el aula o en el centro, se elegiría probablemente estándares absolutos. A la hora de tomar decisiones para una selección donde hay más candidatos que plazas, probablemente se tendrá que utilizar estándares absolutos para poder optar a ser candidato, pero se tendrá que recurrir a estándares normativos para la selección final. Por ejemplo, si se hace una selección de trompetistas para la banda de graduación, se seleccionará sólo entre el 2% que constituye los mejores.

Aún no se ha hablado de cómo establecer los estándares. ¿Cómo se sabe dónde fijar el nivel aceptable de rendimiento? ¿A quién se considera competente? ¿Dónde está el punto que divide lo que es poco satisfactorio de lo satisfactorio? Las evaluaciones trascendentales, como pueden ser las notas de graduación, recurren a procedimientos formales para establecer estándares. Entre estos se puede incluir la utilización de un tribunal evaluador, al que se le ha proporcionado información de referente normativo y de criterios, para definir que requisitos se tienen que cumplir para ser aprobado. En una evaluación a nivel distrito o centro, la calificación de aprobado o los descriptores para un rendimiento deficiente y sobresaliente se determinan por consenso entre aquellos que están utilizando la evaluación. En el aula, los maestros establecen estándares basados en sus experiencias, sus conocimientos de lo que los alumnos han hecho anteriormente, su familiaridad con las expectativas de alguna disciplina, el rendimiento actual de los alumnos y la finalidad de la evaluación.

Considerar otras opciones: Criterios integrales o analíticos*

Basándonos en la experiencia de la evaluación de expresión escrita directa, ofrecemos dos opciones más para especificar criterios: integral o analítico. Los criterios integrales requieren que los calificadores asignen una sola calificación basada en la calidad global o en un sólo aspecto de la respuesta del alumno. Un baremo analítico requiere que los calificadores pongan calificaciones por separado para los distintos aspectos del trabajo. Los criterios que incorporan varias finalidades son analíticos.

*Quizá el término "Primary Trait Scoring" (Calificación de las Características Principales) les resulte familiar. Cuando los criterios de las Características Principales se concentran en una sola son integrales; cuando hay dos o más características, se convierten en analíticos.

¿Cuál es mejor?

Llegados a este punto, ya se pueden imaginar lo que vamos a contestar: "depende del objetivo de la evaluación". La variedad de resultados de un baremo analítico proporciona una retroalimentación útil sobre los puntos fuertes y débiles del alumno individual y el programa didáctico de la clase. Desafortunadamente, como el rendimiento del alumno en dimensiones distintas de un baremo analítico puede relacionarse de formas muy complejas, quizás los resultados no sean tan claramente diagnósticos como se deseaba. A pesar de que una de las características de un buen baremo analítico, desde una perspectiva de eficacia y medición, es que cada dimensión sea distinta, muchas veces las calificaciones de los sub-baremos están muy interrelacionadas y mal diferenciadas. La investigación de CRESST sobre los baremos de calificación analíticos encontró altas correlaciones entre las calificaciones para la organización global de la redacción y de los párrafos, y entre la calificación para la organización, los argumentos de apoyo y la competencia general. Bajo estas circunstancias, el valor diagnóstico del rendimiento de sub-baremo se ve enormemente reducido.

La calificación integral es normalmente más simple y más rápida que la analítica: un asunto importante cuando se tiene en cuenta el tiempo del maestro. A menos que el objetivo de la evaluación no sea el de proporcionar datos para ayudar a mejorar el programa, una rápida impresión del logro alcanzado podría ser especialmente apropiado para la evaluación del programa, para alumnos que necesitan más ayuda y para asignar las evaluaciones finales.

La utilización simultánea de estrategias analíticas e integrales podría mejorar tanto el valor diagnóstico como su eficacia. Un método que ha surgido a partir de las pruebas de competencia mínima es el de calificar todas las redacciones de forma integral y luego evaluar de forma analítica aquellas redacciones que se puntuaron por debajo de la competencia mínima. Otra estrategia, utilizada en la evaluación del estado de Maine, es la de calificar las redacciones de forma integral, pero anotando las dimensiones analíticas que son particularmente fuertes o débiles en el trabajo individual como un "comentario" genérico sobre el rendimiento.

Las opiniones sobre el valor de estos métodos difieren considerablemente y la investigación continúa. Lo importante no es tanto la denominación correcta de baremos, sino el hecho de que existen una variedad de métodos que pueden ser útiles.

¿Y la evaluación de carpetas de trabajo?

La evaluación de carpetas de trabajo es normalmente la primera estrategia que nos viene a la mente cuando pensamos en evaluaciones alternativas. En algunos sentidos, la evaluación de carpetas de trabajo se aplica equivocadamente a "la evaluación de un cuerpo de trabajo". En otros contextos, la evaluación de carpetas de trabajo es realmente el sistema de evaluación. Las carpetas de trabajo son recopilaciones de los trabajos del alumno que se revisan siguiendo unos criterios

para valorar a un alumno en particular o a un programa. La carpeta o recopilación de trabajos no constituye la evaluación; es simplemente un recipiente de los trabajos (redacciones, cintas de vídeo, trabajos artísticos, diarios, etc.) que pueden o no ser evaluados. La "evaluación" de carpetas sólo se da cuando (1) se define una finalidad de la evaluación; (2) se especifican los criterios o métodos para decidir qué se va a incluir en la carpeta, quién lo hace y cuando; y (3) se identifican los criterios que se van a utilizar para evaluar bien la recopilación de trabajos o los trabajos individuales. Decidir lo que se debería incluir es realmente una descripción de la tarea, no un problema de directrices de calificación. Lo que se incluye, quién elige, cuándo se recogen las muestras—éstas son dimensiones de la tarea de evaluación que definen el contexto y los tipos de trabajo que se van a considerar. (Véase el capítulo 7 para más información sobre la evaluación de carpetas).

Existen dos temas relacionados con la selección de las dimensiones de los criterios de calificación para la evaluación de carpetas: (1) ¿Cuáles son los criterios que se utilizan para seleccionar las muestras que se incluyen en la carpeta? y (2) ¿cuáles son los criterios para evaluar la calidad de las muestras? Antes de considerar los criterios para valorar las carpetas, se tendrá que determinar si se debería calificar la carpeta como una entidad o como muestras individuales. En segundo lugar, habrá que decidir cuáles son las dimensiones que reflejan la intención u objetivo de la evaluación. Cuando se examina un cuerpo de trabajo, surgen muchas cuestiones, por ejemplo:

- ¿Se va a valorar el progreso o la mejoría?
- ¿Se va a evaluar el progreso y cómo?
- ¿Cómo se va a comparar, o dar importancia, a las distintas tareas, vídeos, trabajos artísticos, redacciones, diarios u otros en la evaluación?
- ¿Cuál es el papel del alumno en la evaluación? ¿y la aportación de los padres?

Una vez que se hayan solucionado estas cuestiones, definir las dimensiones de los criterios de calificación de carpetas es lo mismo que definir los criterios multidimensionales. Quizá el ejemplo más conocido de los criterios de evaluación de carpetas es el de la carpeta de matemáticas de Vermont, que se resume en el cuadro 5.10. Un cuerpo de trabajos de matemáticas se evalúa en dos dimensiones principales, la resolución de problemas y la habilidad comunicativa. Dentro de cada dimensión, varias subdimensiones definen con más precisión cada una de las destrezas superiores. Se califican las subdestrezas teniendo en cuenta las dos dimensiones, resolución de problemas y comunicación. Como se ve, este ejemplo de criterios de evaluación de carpetas se parece a los ejemplos multidimensiones de los cuadros 5.1 y 5.7.

Cuadro 5.10
Formulario de calificación de matemáticas

Alumno: _____
D.N.I.: _____
Escuela: _____
Curso: _____ Fecha: _____
Calificador: _____

	A1 Comprensión de la tarea FUENTES DE EVIDENCIAS • Explicación de la tarea • Adecuación del método • Respuesta correcta que demuestra comprensión	A2 Cómo—calidad de métodos/procedimientos FUENTES DE EVIDENCIAS • Demostraciones • Descripciones orales o escritas • Borradores, trabajos en sucio, esbozos	A3 Por qué—decisiones tomadas a lo largo del proceso FUENTES DE EVIDENCIAS • Cambios de método • Explicaciones orales o escritas • Validez de la solución • Demostración
ENTRADA 1 Título: _____ P Prob I Invest A Apl O Otro	☐	☐	☐
ENTRADA 2 Título: _____ P Prob I Invest A Apl O Otro	☐	☐	☐
ENTRADA 3 Título: _____ P Prob I Invest A Apl O Otro	☐	☐	☐
ENTRADA 4 Título: _____ P Prob I Invest A Apl O Otro	☐	☐	☐
ENTRADA 5 Título: _____ P Prob I Invest A Apl O Otro	☐	☐	☐
ENTRADA 6 Título: _____ P Prob I Invest A Apl O Otro	☐	☐	☐
ENTRADA 7 Título: _____ P Prob I Invest A Apl O Otro	☐	☐	☐
CALIFICACIONES CLAVE →	COMPRENSIÓN DE LA TAREA CALIFICACIÓN FINAL 1 No comprendida 2 Comprensión parcial 3 Comprendida 4 Generalizada aplicada extendida	CÓMO—CALIDAD DE LOS MÉTODOS/PROCEDIMIENTOS CALIFICACIÓN FINAL 1 Método/procedimiento inapropiado o no viable 2 Método/procedimiento apropiado en algún momento 3 Método/procedimiento viable 4 Método/procedimiento eficaz o sofisticado	POR QUÉ DECISIONES A LO LARGO DEL PROCESO CALIFICACIÓN FINAL 1 No hay evidencias de una toma de decisión razonada 2 Posible toma de decisión razonada 3 Decisiones/ajustes razonados inferidos con certeza 4 Decisiones/ajustes razonados demostrados/explicados
Comentarios:			

Cuadro 5.10 *(continuación)*

A4 Qué—resultados de actividades FUENTES DE EVIDENCIAS • Soluciones • Extensiones—observaciones, conexiones, aplicaciones, síntesis, generalizaciones, abstracciones	B1 Lenguaje matemático FUENTES DE EVIDENCIAS • Terminología • Anotaciones	B2 Representaciones matemáticas FUENTES DE EVIDENCIAS • Gráficas, tablas, dibujos • Modelos • Diagramas • Manipulaciones	B3 Claridad de la presentación FUENTES DE EVIDENCIAS • Forma de describir/escribir la(s) soluciones • Trabajo escrito • Entrevistas con el maestro y observaciones • Errores de lógica • Comentarios del alumno/respuestas • Autoevaluación del alumno	TOTALES DE CONTENIDO [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible]
☐	☐	☐	☐	
☐	☐	☐	☐	
☐	☐	☐	☐	
☐	☐	☐	☐	
☐	☐	☐	☐	
☐	☐	☐	☐	
☐	☐	☐	☐	
QUÉ—RESULTADOS DE LAS ACTIVIDADES **CALIFICACIÓN FINAL** 1 Solución sin más 2 Solución con observaciones 3 Solución con conexiones o aplicaciones 4 Solución con generalización de síntesis o abstracción	**LENGUAJE MATEMÁTICO** **CALIFICACIÓN FINAL** 1 Uso inexistente o inapropiado del lenguaje matemático 2 Uso apropiado del lenguaje matemático en determinados momentos 3 Uso apropiado del lenguaje la mayoría del tiempo 4 Uso del lenguaje matemático apropiado preciso y elegante	**REPRESENTACIONES MATEMÁTICAS** **CALIFICACIÓN FINAL** 1 Ningún uso de representaciones matemáticas 2 Uso de representación(es) matemática(s) 3 Uso correcto y apropiado de representación(es) matemática(s) 4 Uso inteligente de representación(es) matemática(s)	**CLARIDAD DE LA PRESENTACIÓN** **CALIFICACIÓN FINAL** 1 Poco clara (o es organizado incompleto: faltan detalles) 2 Parcialmente clara 3 En gran parte clara 4 Clara (o, bien organizada completa detallada)	CARACTERÍSTICAS DE LA TAREA COMENTARIO SOBRE ATRIBUTOS CON PESO Motivación Flexibilidad Conocimiento Reflexión Confianza Perseverancia Curiosidad/Invención Valor matemático

BEST COPY AVAILABLE

Diseñar y evaluar los criterios de evaluación

Comienzo del proceso del diseño

El proceso del diseño de nuestros propios criterios no constituye un problema:

- Investigar cómo define la disciplina evaluada un rendimiento de calidad.
- Recopilar rúbricas ejemplares para evaluar la expresión escrita, oral, artística, etc., como modelos que se pueden adaptar a los objetivos.
- Recopilar muestras de trabajos de alumnos y expertos que reflejen la gama de rendimiento que va desde ineficaz a muy eficaz.
- Hablar con otros sobre las características de estos modelos que distingan a los eficaces de los ineficaces.
- Escribir descriptores para las características importantes.
- Recopilar otra muestra de trabajos de alumnos.
- Someter a prueba los criterios con el fin de comprobar si ayudan a valorar con precisión a los alumnos.
- Revisar los criterios.
- Intentar de nuevo hasta que la calificación de la rúbrica llegue a captar la "calidad" del trabajo.

Probablemente se apreciará lo reincidente que es este proceso de diseño. Las ideas iniciales sobre los aspectos importantes y calificables del rendimiento del alumno se perfeccionan con la práctica. Los criterios pueden centrarse en el proceso—cómo un alumno aborda y soluciona un problema—y a su vez pueden enfocarse en el producto o los resultados.

Por ejemplo, podemos referirnos al proceso de diseño de los criterios en el cuadro 5.5 (Baker, Aschbacher, Niemi y Sato 1992). CRESST diseñó su rúbrica para calificar el grado de comprensión del contenido en la asignatura de historia por medio de la recopilación y examinación de las diferencias que había entre las redacciones escritas por expertos en historia (profesores universitarios y estudiantes de posgrado de historia) y aquéllas escritas por novatos (alumnos de instituto). Los investigadores de CRESST buscaron dimensiones que parecían diferenciar el rendimiento de estos dos grupos. En un número de áreas de programas de estudio, los investigadores observaron diferencias entre los alumnos y los expertos en la aplicación de conocimientos previos, la utilización de conceptos y principios organizadores y conceptos equivocados. Estos rasgos definieron el primer borrador de los criterios de calificación. Seguidamente, se probaron estos criterios en muestras de trabajos de alumnos y fueron aclarados y refinados para asegurar que los baremos estuvieran bien definidos, fueran apropiados para la variedad de respuestas de alumnos que esperaban encontrarse.

y que permitirían a los maestros u otros calificadores distinguir entre las redacciones que merecían notas contiguas en la baremo.

Mientras se realiza la tarea de diseño de los criterios, no se debe olvidar el aprovechamiento del trabajo de otros. Muchas veces se puede importar o modificar criterios procedentes de programas de evaluación estatales y locales, expertos en programas de estudio o colegas que han luchado con problemas de evaluación similares. La literatura de investigación sobre la evaluación alternativa también proporciona ejemplos de evaluaciones alternativas piloto parecidas a la que se muestra en el cuadro 5.4, que se puede adaptar para su uso en el aula. También existe una literatura reducida, pero en expansión, sobre la naturaleza de la habilidad en varias disciplinas, como por ejemplo la manera en que un historiador lee y utiliza fuentes originales.

Evaluar los criterios

Los criterios para valorar los trabajos del alumno afectan las decisiones que se toman eventualmente sobre programas y alumnos. Sin tomar en cuenta si se están diseñando los propios criterios o si se están utilizando aquellos proporcionados por otros, es importante revisar la calidad de las directrices de calificación. Terminamos este capítulo con la propuesta de un conjunto de "criterios para los criterios"—una lista de control que se puede utilizar para valorar la calidad de los criterios de calificación que bien se toman prestados o se diseñan. Una propuesta de estos criterios aparece en el cuadro 5.11.

Ahora vamos a ver un conjunto de dimensiones para valorar los criterios de cada uno.

Relación con las finalidades más importantes

Como mínimo, los criterios para evaluar el rendimiento del alumno tienen que responder a todos los objetivos que se está intentando medir. Por ejemplo, los criterios para evaluar representaciones dramáticas de los alumnos deberían incluir todos los aspectos importantes del arte dramático y la expresión artística que se quiere evaluar, y no otros. Si la originalidad y la presentación lógica forman parte de las objetivos deseados, se querrá incluir baremos para evaluar estos aspectos del trabajo del alumno. Si no constituyeran un objetivo importante, se deben omitir.

Sensibilidad al objetivo

¿Cuáles son las decisiones educativas que se tomarán a partir de la evaluación? La respuesta a esta pregunta debería guiar las decisiones sobre si se debe utilizar una lista de control o baremo de calificación, qué número de baremos, qué características, qué tipos de baremo, etcétera. ¿Se necesita una visión global e

integral del logro del alumno o una visión analítica que proporcione información sobre varios aspectos específicos del rendimiento del alumno? ¿Se necesita esta información en forma numérica para facilitar su interpretación y suma en perjuicio de los detalles, o se necesita la riqueza de una descripción cualitativa, o quizás las dos cosas?

Cuadro 5.11
¿Cómo evaluar los criterios de calificación?

- ❒ Todas los objetivos importantes responden a los criterios
- ❒ La estrategia de calificación corresponde a la finalidad de la decisión: integral para una visión global y evaluadora; analítica para una visión diagnóstica.
- ❒ El baremo de calificación proporciona calificaciones útiles y fáciles de interpretar.
- ❒ Los criterios emplean referencias concretas, un lenguaje claro, comprensible tanto para alumnos, como para padres y otros maestros.
- ❒ Los criterios reflejan los conceptos actuales de "excelencia" aceptados en el campo.
- ❒ Los criterios han sido revisados para eliminar prejuicios de desarrollo, étnicos y de sexo.
- ❒ Los criterios reflejan objetivos que se pueden enseñar.
- ❒ Los criterios se limitan a un número de dimensiones viable.
- ❒ Los criterios se pueden aplicar a otras tareas similares o a una área de rendimiento superior.

Significativos, claros y creíbles

Los criterios que se emplean para valorar un rendimiento tienen que ser significativos para los alumnos, padres, calificadores maestros, administradores, coordinadores y el público en general. Si los criterios no son creíbles probablemente se ignorarán los resultados o se utilizarán incorrectamente. Ejemplos de trabajos de alumnos que ilustran las características de los criterios pueden ayudar a otros a entenderlos. Incluir a otras personas en el diseño de los criterios aumenta su credibilidad.

81

Puesto que uno de los principios de la evaluación del rendimiento reside en criterios públicos y hablados, los criterios tienen que tener sentido para los alumnos para que los puedan aplicar fácilmente a su propio trabajo y así convertirse en alumnos autónomos. Aunque las opiniones sobre el rendimiento del alumno tienden a ser subjetivas por naturaleza, son más fiables y creíbles cuando dependen menos de un alto nivel de inferencia y más de características observables y concretas.

Justas y objetivas

Las tareas de evaluación no sólo deben ser justas, sino que también deben serlo los criterios que se utilizan para definir la excelencia. Prejuicios no reconocidos pueden infiltrarse en las definiciones de las características, las especificaciones sobre qué tipo de rendimiento corresponde a cada uno de los puntos del baremo y la aplicación de aquellos criterios a trabajos individuales de alumnos. Cuando se desea que los criterios tengan un valor diagnóstico, éstos deben ser sensibles a la enseñanza y a las oportunidades que tienen los alumnos de aprender las destrezas que se van a evaluar. Por el contrario, no se quiere que reflejen variables sobre las cuales los educadores no tienen poder alguno, como por ejemplo la cultura, sexo, o entorno socioeconómico de un niño.

Viables

Existen muchas razones que limitan el número y la complejidad de las dimensiones de rendimiento que se van a valorar. En primer lugar, el tiempo, el esfuerzo y el dinero disponible para valorar el rendimiento siempre están limitados, a veces gravemente. En segundo lugar, los calificadores suelen tener problemas a la hora de tener en cuenta demasiados aspectos de un trabajo a la vez. En nuestra experiencia en CRESST, los calificadores terminaron frustrados cuando les pedimos que utilizaran más de seis o siete baremos para evaluar redacciones de alumnos. Se convirtió en una tarea pesada y en un proceso menos fiable. En tercer lugar, probablemente los alumnos encontrarán difícil tratar demasiados aspectos a la vez de su trabajo. Finalmente, los administradores y coordinadores normalmente necesitan información de la forma más breve posible. Las calificaciones por separado para un gran número de características complejas quizá vuelva aún más difícil la eficaz utilización de los resultados.

Generalizables

Aunque se reconoce que los criterios para un rendimiento están muy relacionados con las nociones basadas en disciplinas de lo que constituye la excelencia, la calificación puede ser más eficaz cuando un sólo conjunto de criterios "genéricos"

puede servir para temas, tareas o disciplinas múltiples. Por ejemplo, se podría establecer un conjunto común de criterios para evaluar la comprensión del alumno de conceptos científicos por medio de diarios, experimentos prácticos, simulacros en computadora y presentaciones orales. ¿Se podría también utilizar un conjunto común de criterios para evaluar redacciones de alumnos en las asignaturas de ciencias sociales, naturales y matemáticas? Aunque estas situaciones parezcan diferentes, es posible incluir criterios genéricos para algunos objetivos. Si pudiéramos conceptualizar la excelencia de manera constante mediante métodos de evaluación y disciplinas, nuestros criterios podrían tener un impacto más fuerte en el aprendizaje y la enseñanza. Nuestro ejemplo de la rúbrica de historia-ciencias sociales de CRESST (cuadro 5.5) que también se ha aplicado a las ciencias naturales y a las económicas, muestra una estrategia para desarrollar criterios multidisciplinarias. Como cualquier criterio bueno, estas dimensiones propuestas están sujetas a una revisión y refinamiento.

Referencias bibliográficas

Baker, E.L., P.R. Aschbacher, D. Niemi y E. Sato. (1992). *CRESST Performance Assessment Models: Assessing Content Area Explanations*. Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Baxter, G., R.J. Shavelson, S. Goldman y J. Pine. (Primavera 1992). *Journal of Educational Measurement* 29, 1: 1-17.

Jones, E. y J.M. Roberts. (1990). *Profile of Developmental Outcomes for Kindergarten Literacy and Numeracy Skills*. Soledad, Calif.: San Vicente School District.

Vermont Department of Education. (1992). *Looking Beyond "the Answer": The Report of Vermont's Portfolio Assessment Program. Pilot Year 1990-91*. Montpellier: Vermont Department of Education.

6

Asegurar una calificación fiable

Una de las características fundamentales de la evaluación del rendimiento es la de depender del juicio humano. Como diría un abogado, dos personas que son testigos de un mismo suceso, o que leen un mismo documento, muchas veces perciben o interpretan de forma distinta. De la misma manera, aquellos que ven la misma conducta en ocasiones distintas pueden llegar a opinar de forma diferente sobre esa conducta. El usuario o coordinador de evaluaciones alternativas debe tratar de minimizar estas diferencias: de no ser así, las valoraciones no serán justas, constantes o válidas. Los buenos procedimientos de calificación fomentan este proceso.

Entender la importancia de la fiabilidad y la constancia

La razón más clara para una calificación constante es la equidad. Las valoraciones sobre el rendimiento del alumno no pueden ser caprichosas si queremos que sean significativas. Es preciso tener la seguridad de que la nota o valoración fue el resultado del propio rendimiento y no de un aspecto superficial del producto o del

contexto de la calificación. ¿Se vio la nota de Yuki afectada injustamente por sus faltas de ortografía? ¿Sacó Mark una mejor (o peor) nota porque se corrigió su proyecto casi al final, cuando el maestro ya estaba cansado? ¿De qué manera se vio afectada la nota de Jamal por el hecho de que otro maestro participara en el proceso de corrección? ¿Y Corinne, suspendió el examen de expresión escrita porque los evaluadores de este año fueron más exigentes que los del año pasado?

La falta de constancia es especialmente problemática cuando los resultados influyen en decisiones importantes sobre alumnos o programas. ¿Cuál es la nota que merece Denisha? ¿Debemos dejar que Marta se matricule en la clase de nivel avanzado de lengua inglesa a pesar de sus bajas notas en las pruebas estandarizadas? ¿Debería seguir funcionando el nuevo programa de matemáticas del centro? Incluso cuando los resultados de una evaluación aislada no conllevan decisiones trascendentales, la falta de constancia conduce a una calificación inexacta. Para ser más concisos: una calificación no constante indica que las notas tienen poco valor. Si una "A" no representa de manera constante un rendimiento sobresaliente, ¿entonces qué quiere decir? ¿El mejor de la clase? ¿El mejor de un grupo con un bajo nivel? ¿Una mejora? Si un rendimiento o un proyecto recibe una calificación distinta de diferentes calificadores, ¿qué significa cada una de ellas? ¿Cuál es la más exacta? Si se utilizan los criterios de manera diferente según el período empleado en la calificación, ¿qué significa la calificación final? ¿Qué indica la puntuación de un alumno en particular?

Lograr la constancia

Una calificación justa y significativa requiere de un juicio apropiado y constante. ¿Cómo se puede evitar la subjetividad caprichosa? Como establecimos en el capítulo 5, tener criterios bien definidos y justificables para valorar el rendimiento de un alumno contribuye enormemente en conseguir un sistema de calificación constante, pero existen otras condiciones que hay que reunir para poder asegurar esta constancia. En primer lugar, aquellos que van a valorar—ustedes, sus colegas, el departamento estatal de educación—deben entender los criterios de igual manera. La base de la constancia de la calificación es el consenso entre los calificadores sobre lo que significan los criterios y cómo se van a emplear. En segundo lugar, se necesita un sistema para controlar la constancia de las calificaciones durante el período en el que se evalúa el rendimiento. Esta constancia tiene varias facetas. Dos o más calificadores que evalúan el mismo rendimiento deben llegar a un acuerdo general. Un calificador debe calificar un determinado rendimiento de la misma manera sin importar cuándo fue observado, ya sea al principio del día, al mediodía o casi al final. Los calificadores deben calificar de forma parecida los mismos rendimientos en distintas ocasiones. Y los mismos rendimientos evaluados en dos ocasiones distintas por dos grupos distintos de calificadores deben igualmente calificarse de manera similar. Si las calificaciones se van a emplear para tomar decisiones trascendentales como son pasar de un curso a otro, la graduación o la asignación a clases de atención especial, se debe documentar de manera formal/oficial las evidencias de constancia en la calificación.

Las ventajas del desarrollo profesional

El proceso que utilizan los calificadores para aprender a utilizar los criterios de calificación de manera constante puede proporcionar una valiosa oportunidad para el desarrollo profesional. La formación de los calificadores ayuda a los maestros a llegar a una definición consensual sobre los aspectos clave del rendimiento escolar. Esto puede conducir tanto a una nueva prioritización de los objetivos didácticos como a información acerca de los puntos fuertes y débiles del rendimiento de los alumnos. El proceso de calificar puede proporcionar un modelo de evaluación en el aula y puede fomentar una mayor colaboración entre los maestros en la evaluación de objetivos escolares.

Para poder sacar provecho de la constancia y del crecimiento profesional, se requerirá procedimientos de formación buenos y un proceso de calificación cuidadosamente estructurado. Este capítulo describe las consideraciones más importantes cuando se va a diseñar y poner en práctica un procedimiento de calificación válido. Aunque el proceso que describimos tiene sus orígenes en evaluaciones formales y trascendentales a nivel distrito y estatal, se deberá recordar que una calificación constante se puede aplicar a todo tipo de evaluación, sean calificaciones durante el curso escolar o pruebas de selección para ingreso a la universidad. Las decisiones que se toman sobre un estudiante no pueden ser válidas si no se basan en información fiable.

La formación de calificadores: Un requisito para la calificación constante

Hay varias formas de conseguir la constancia. Nuestro planteamiento enfatiza la formación de calificadores a un estándar común, ya que esto es eficaz y proporciona a los maestros información de valor didáctico. Otros tipos de planteamientos dedican menos atención a la formación de calificadores y a la construcción de consensos, y utilizan múltiples opiniones sobre el trabajo del alumno para lograr un resultado parecido. Como se puede imaginar, el método que se elige depende del objetivo de la evaluación y de los recursos disponibles.

Durante la formación de calificadores, éstos aprenden lo que significan los criterios de calificación, qué aspectos del rendimiento ha de evaluar cada uno y lo que cada uno de los puntos del baremo representa. Durante el período de formación es preciso asegurar que los calificadores apliquen de forma constante los criterios a una amplia gama de muestras de trabajos de alumnos. También es ahora cuando los calificadores aprenden cómo documentar sus calificaciones.

Manuales de formación

Los manuales formales de calificación pueden ser de gran utilidad tanto durante la formación como posteriormente. Para las evaluaciones a gran escala, como por ejemplo los programas de evaluación anuales a nivel distrito o estatal, un manual

de calificación proporciona una "memoria institucional" de los procedimientos de evaluación y sirve como útil referencia para la interpretación de las calificaciones. Para evaluaciones trascendentales en el aula, como por ejemplo pruebas de selectividad para ingreso a cursos de alto nivel (Advanced Placement), o una prueba de agilidad algebraica, los manuales de calificación pueden ser de utilidad en las conversaciones con padres o alumnos que quieren saber cómo lograr o mejorar determinadas calificaciones. Las guías típicas de calificación incluyen:

- criterios de calificación explicados totalmente;
- ejemplos o modelos que ilustran cada calificación del baremo;
- una versión abreviada de una página de los criterios o referencias durante la calificación actual y
- un modelo de formulario para anotar las calificaciones.

Si se desea, se puede revisar manuales de evaluación procedentes de varias fuentes antes de diseñar un curso propio para la formación de calificadores. Si se está interesado en la descripción detallada del proceso de formación de calificadores, un manual completo de calificación diseñado por el Riverside Publishing Company aparece en *Educational Performance Assessment*, editado por Fred Finch (1991). Los departamentos estatales de educación también son fuentes de manuales de formación que han sido publicados.

Procedimientos de formación

La formación de calificadores está diseñada para crear un grado de comprensión consensual de los criterios de calificación, proporcionar una práctica extensiva al calificar y, en el caso de una evaluación trascendental, proporcionar niveles aceptables de constancia de calificación (la fiabilidad). Durante la formación de los calificadores, las sesiones prácticas de calificación proporcionan a éstos una retroalimentación inmediata y sustancial acerca de sus decisiones así como muchas oportunidades para formular preguntas. Los calificadores también llegan a entender que su trabajo consiste en hacer valoraciones basándose en la rúbrica de calificación, no en revisar o criticar la rúbrica y luego seguir sus propias idiosincrasias. De no entender estos principios, todo un proyecto de evaluación puede venirse abajo.

Una típica sesión de formación incluye:

- **Orientación hacia la tarea de evaluación.** Los calificadores reciben una descripción del contexto de la evaluación, para qué se van a utilizar los resultados, quién los va a utilizar, qué instrucciones y pautas recibieron los alumnos, y cómo es que la guía de calificación hace operativo los objetivos o procesos deseados. Es habitual pedir a los calificadores que realicen la prueba ellos mismos como medio de orientación para la tarea de calificación.

- **Aclaración de los criterios de calificación.** En esta fase de la formación, los calificadores entran en un proceso de debate amplio. Tanto las dimensiones de los criterios como los valores del baremo se definen y se proporciona toda una serie de modelos que ejemplifican cada uno de ellos. Los debates normalmente abarcan desde lo que son los juicios más simples, como por ejemplo lo que constituye una muestra de un rendimiento alto, medio o bajo, hasta diferenciaciones más difíciles que son necesarias para las calificaciones numéricas.
- **Prácticas de calificación.** Este es el elemento más importante del proceso de formación de calificadores. Al principio se puntúan exámenes muestra uno por uno para luego hablar de ellos. A la vez que los calificadores van adquiriendo más soltura con los baremos de calificación, tienen la oportunidad de tomar decisiones más difíciles en cuanto a evaluaciones problemáticas (atípicas) o dudosas.
- **Revisión protocolaria.** Durante la discusión y las prácticas de calificación, los calificadores normalmente establecen algunas reglas para lidiar con aspectos de valoración inesperados que presenta un determinado conjunto de exámenes y que no entran dentro del baremo de calificación. Por ejemplo, cuando casi todos los alumnos han malinterpretado del mismo modo una pregunta, en lugar de calificar todas las respuestas como "irrelevante" o "inaceptable", los calificadores pueden decidir dar calificaciones que se basan en la interpretación que el alumno ha hecho de la tarea. O, si hay que calificar muchas características diferentes, los calificadores pueden decidir que determinados calificadores se dediquen a la calificación de algunas de las características, en lugar de estar todos calificando cada examen en todas sus dimensiones.
- **Anotación de las calificaciones.** En todas las evaluaciones, se deberá anotar de alguna manera las calificaciones de los alumnos, en listas o en las actas de clase, curso, o centro. La formación de los calificadores incluye el formato para la anotación de las calificaciones y todos los procedimientos especiales para calcular las notas de los alumnos, como por ejemplo sacar la media o el total por dimensiones.
- **Documentación de la fiabilidad de los calificadores.** La formación de los calificadores termina cuando hay un acuerdo de que todos los calificadores han llegado a un nivel aceptable de constancia, normalmente cuando la calificación de las muestras difiere tan sólo en un punto. Para poder decidir cuándo los calificadores están debidamente capacitados, se llevan a cabo pruebas de fiabilidad durante la formación. El cuadro 6.1 presenta un ejemplo de cómo comprobar la constancia de los calificadores utilizando el sistema de acuerdo por porcentaje.
- **Factores a considerar en la programación.** ¿Cuánto tiempo llevará la formación de calificadores hasta que lleguen a un nivel aceptable antes de permitirles corregir trabajos de alumnos? Esto dependerá de:

  — La experiencia de los calificadores.

— Su familiarización con los criterios de calificación.
— La rapidez con la que los calificadores consigan llegar a un consenso sobre lo que significan los criterios.
— La complejidad de los criterios de calificación y la calidad del trabajo que se ha de evaluar—siendo los trabajos dudosos los más difíciles de evaluar con rapidez.

Hemos comprobado que se requiere de tres a cuatro horas para preparar a los calificadores en la utilización de un baremo integral o analítico simple (de dos a cuatro características). Los baremos más complicados requieren de casi todo un día de preparación.

Cuadro 6.1
Cálculo del acuerdo entre calificadores (Tres calificadores para dos exámenes)

	¿Está el calificador totalmente de acuerdo con la calificación establecida?			¿Está el calificador de acuerdo con la calificación establecida en más/menos un punto?		
Calificador	Examen Nº 1	Examen Nº 2	Media de acuerdo entre los calificadores	Examen Nº 1	Examen Nº 2	Media de acuerdo entre los calificadores
Linda	sí	no	50%	sí	no	50%
Robert	no	no	0%	sí	sí	100%
Ellia	sí	sí	100%	sí	sí	100%
Total	67% = sí	33% = sí	50%	100% = sí	67% = sí	83%

El cuadro 6.1 ilustra el caso de tres calificadores a los que se les pidió que calificaran dos exámenes tipo después de alguna formación previa. Según los resultados que muestra el cuadro, Linda está de acuerdo con la calificación modelo para el primer examen, pero no con la del segundo; de hecho, referente al segundo examen demuestra más de un punto de desacuerdo con la calificación establecida. Robert no está totalmente de acuerdo con las calificaciones modelo ni para el primer examen, ni para el segundo, pero está de acuerdo con más/menos un punto de diferencia con la calificación de ambos exámenes. Ellia está de acuerdo con todas y está preparada para calificar trabajos de alumnos. Robert y Linda probablemente necesitan un poco más de formación. El segundo examen causa más problemas a los calificadores que el primero, así que la formación adicional debería centrarse en la capacidad de distinguir la calificación establecida de las calificaciones contiguas. A la hora de describir estos resultados se podría decir, "Como promedio, los calificadores llegaron a un total acuerdo con las calificaciones modelo en un 50% de las veces, y llegaron a un acuerdo de más/menos un punto en un 83% de las veces."

El cansancio de los calificadores es un factor importante en la calificación; consideramos que un día completo de trabajo equivale a una sesión de seis horas. También se deberá programar el tiempo para volver a capacitar a los calificadores o para refrescarles la memoria al principio de cada nuevo día que se dedica a la calificación y, por supuesto, para cualquier cambio de temas o tareas que emplean los mismos criterios de calificación. En una evaluación trascendental, el volver a capacitar al calificador tiene normalmente lugar después de un buen descanso, como por ejemplo tras el almuerzo.

Temas relacionados con los exámenes tipo utilizados durante la formación

Dado que la formación de los calificadores constituye un ensayo para la calificación real, es necesario anticipar tantas fuentes de desacuerdo entre calificadores como sea posible antes de su formación y sacar ejemplos de los exámenes utilizados que faciliten el desacuerdo y la discusión. Por ejemplo, las construcciones sintácticas utilizadas por hablantes no naturales de inglés hacen surgir temas relacionados con la relación entre el contenido y los objetivos de la comunicación. También se deberían tratar asuntos relacionados con la letra de los alumnos y la legibilidad o cuestiones de calidad estética en las artes plásticas o del escenario. Finalmente, debemos asegurarnos de que los exámenes tipo que se seleccionan para la formación reflejen no sólo cada punto del baremo a utilizar, sino también toda la gama del rendimiento del alumno que es probable que se encuentre en la evaluación. La tendencia humana natural es la de calificar normativamente. Las mejores muestras de trabajo de un conjunto de exámenes relativamente flojos pueden tener una calificación más alta de la que recibirían si formaran parte de un grupo de exámenes relativamente fuertes. De igual manera, también podría ocurrir lo contrario. Se deberá hablar sobre esta tendencia durante la formación de los calificadores y acompañarla de ejemplos para que los criterios de calificación mantengan el mismo significado en los diferentes conjuntos de exámenes y durante todas las distintas sesiones de calificación.

Obtención de exámenes tipo

Dado que se necesita una amplia selección de ejemplos de trabajos para guiar a los calificadores, se debe recopilar muestras de un grupo heterogéneo de alumnos. Se deberá seleccionar muestras de un trabajo de campo, de una evaluación anterior o de la evaluación actual. En la selección de exámenes apropiados para la formación y el control, un grupo de "expertos"—maestros de los cursos y asignaturas implicadas en la evaluación que están familiarizados con los criterios de la evaluación—pueden ser de gran ayuda. Estos maestros pueden seleccionar ejemplos que ilustren todas las posibles respuestas, desde la más clara a la más dudosa, para cada uno de los puntos del baremo de manera que los calificadores

estén preparados para enfrentarse a todas las posibles situaciones. Si en la evaluación se utilizan varias preguntas o tareas, se necesitarían ejemplos que ilustren cada una de ellas. Si se está utilizando baremos relacionados con la edad en todos los cursos, se necesitarían ejemplos que ilustren los niveles de las distintas edades. También es útil preparar comentarios escritos que expliquen cómo los aspectos específicos de cada examen reflejan los criterios de una nota en particular. El grupo de expertos podría más adelante identificar muestras que serán utilizadas para (1) discusiones en la sesión de formación, (2) la práctica y (3) la comprobación de la constancia.

Temas relacionados con la documentación de notas

Se deberá proporcionar a los calificadores un método para la documentación de las notas de los alumnos. En la propia aula, las notas quizás se anoten simplemente en la parte superior del examen del alumno y luego en la lista de clase. Algunos maestros utilizan los criterios de calificación como retroalimentación para alumnos. Estos maestros marcan las áreas deficientes o anotan los puntos fuertes consultando los descriptores de la guía. El mismo proceso puede utilizarse para crear un perfil de la clase en una guía maestra de calificación.

En situaciones más formales de evaluación, las libretas de notas se convierten en documentos públicos y se utilizan para proporcionar retroalimentación para maestros y otros. Los analistas de datos también las utilizan para calcular las estadísticas de exámenes. En estas circunstancias, se da a los calificadores documentos informatizados para rellenar casillas y otra información importante como por ejemplo centro, distrito y números de identificación de los calificadores, así como los códigos para el tema o tarea y fecha. Cuando hay dos o más calificadores calificando los trabajos de los alumnos, habrá que recordarles que no indiquen notas, comentarios o correcciones en el papel de examen. No es deseable que una calificación posterior se vea influida por estos comentarios.

Temas relacionados con la fiabilidad

El objetivo de la formación de los calificadores es establecer procedimientos de calificación constantes y fiables. Por consiguiente, se deberá incorporar un método para determinar si los calificadores son constantes durante el período de formación. Existen muchas estrategias para comprobar la fiabilidad de los calificadores. Un método empleado frecuentemente es el de preparar y calificar con antelación un conjunto de unos diez exámenes de "control de fiabilidad" que representen toda la gama del posible rendimiento del alumno. A continuación se pide a los calificadores que califiquen este mismo conjunto para comparar sus calificaciones con las de los otros evaluadores con mayor experiencia. Un nivel de acuerdo razonable tanto con las decisiones de los expertos como con las de entre ellos mismos sugiere que los calificadores están preparados para calificar el trabajo real de alumnos.

¿Qué es lo se debe entender por acuerdo razonable? Se puede exigir que los calificadores lleguen a un acuerdo total antes de considerarles fiables, o se puede utilizar la regla menos rigurosa de "más o menos uno", que es bastante común y que establece que los calificadores están "de acuerdo" si sólo difieren en un punto "más o menos". Por ejemplo, si la calificación de una determinada muestra de control de fiabilidad es un "3", se considera que aquellos que dieron una calificación de "2", "3" ó "4" están preparados.

Sin tomar en cuenta el nivel de acuerdo deseado que se elija, cuando se forma a los calificadores, el objetivo es que aprendan a utilizar los criterios de calificación tal y como se pretendía, y no con un punto de diferencia. Cuando un calificador tiene dificultades en aplicar los criterios tal y como se pretende, se debería dedicar algún tiempo durante la formación a la discusión de los exámenes que se utilizan para practicar, de los criterios y reglas que se han de seguir para aplicarlos con el fin de que el calificador alcance un nivel aceptable de constancia. Sin embargo, algunos calificadores quizás no puedan ajustar sus criterios internos a las guías de calificación. Estos calificadores que no logran adaptarse deberían ser dados de baja o asignados a otras tareas durante la sesión de calificación.

Además de establecer qué diferencia se va a permitir entre los calificadores para lograr la constancia, también se deberá decidir el número de veces que deben conseguir estar de acuerdo. Si lo que se requiere es un acuerdo total, algo difícil de obtener, el criterio de fiabilidad podría ser menos riguroso que si se utiliza la regla de "más o menos uno". En CRESST, normalmente pedimos que los calificadores alcancen un acuerdo con los expertos al menos en un 90 por ciento de las veces para cada dimensión de la calificación cuando se utiliza la pauta de "un punto arriba o abajo". La pauta para el acuerdo total puede reducirse a un 75/80 por ciento bajo condiciones más estrictas. El porcentaje actual de acuerdo varía según el objetivo de la evaluación y la trascendencia de la misma.

Sin tomar en cuenta la definición de lo que es "el acuerdo entre calificadores", el objetivo de los controles de fiabilidad es el de asegurar que las calificaciones de los alumnos no sean el resultado de un juicio caprichoso, uno de los argumentos más citados en contra de la evaluación de rendimiento. Consideremos el conocido estudio llevado a cabo por Paul Deidrich (1963) en el Servicio de Evaluación Educativa (Educational Testing Service) en el que una misma redacción recibió todas las posibles calificaciones de un grupo de calificadores. Lo que la mayoría no recuerda de este estudio es que se obtuvo niveles aceptables de acuerdo entre los calificadores cuando los evaluadores (1) procedían de la misma disciplina, (2) utilizaban criterios de calificación explícitos y (3) habían participado en una sesión de formación previa.

Asegurar juicios equitativos durante una sesión de calificación real

Mantener la constancia

La documentación de la constancia de los calificadores durante su formación es simplemente el primer paso hacia la creación de un procedimiento de calificación justo y equitativo. Puesto que el objetivo de la formación de los calificadores es el de fomentar la constancia entre ellos, también habrá que controlar y vigilar los patrones entre la calificación de los calificadores durante el mismo proceso de calificación. La investigación demuestra que los calificadores tienen tendencia a alejarse de los criterios formales acercándose a sus propias opiniones más idiosincrásicas (Quellmalz y Burry 1983). Los juicios y expectativas humanas se ven afectadas no sólo por los estándares formales, como por ejemplo los criterios de calificación, sino también por su experiencia previa y la gama de rendimiento que se está evaluando en ese momento. Si todo el conjunto de rendimiento parece ser relativamente "deficiente" según los criterios que apuntan al objetivo, los calificadores tienden a bajar los criterios para poder dar notas más altas a los exámenes que están entre lo "mejor de lo peor". Como maestro, uno quizá sea consciente de que los estándares y expectativas que tenemos puestos en los alumnos cambian durante el proceso de calificación. Hasta cierto punto modificamos nuestras ideas después de ver varios trabajos de los alumnos. Por esta razón, las sesiones de formación deben incluir un gran número de exámenes y la completa gama que se podría encontrar durante el propio proceso de calificación.

Para una evaluación realizada en el aula, se puede comprobar la constancia deteniéndose en medio del proceso para volver a puntuar algunos de los primeros trabajos de alumnos ya corregidos. Cuando se va a calificar varias dimensiones o temas diferentes, se puede corregir a la vez sólo una dimensión o el trabajo relacionado con un sólo tema, para luego volver a calificar los otros factores. Muchas veces es más rápido calificar todos los exámenes varias veces, una vez por cada dimensión o tema diferente, que corregir exámenes individuales mirando todas las dimensiones a la vez y aplicando criterios múltiples o leyendo distintos tipos de respuesta. La velocidad de calificación aumenta también conforme uno se va familiarizando con los criterios.

Para una evaluación a nivel centro, de mayor escala, o trascendental, es deseable incluir más controles formales de constancia entre los calificadores. Para la calificación de redacciones, a veces esto se lleva a cabo introduciendo exámenes de control ya calificados a intervalos designados entre los exámenes de cada calificador. El director de calificación comprobará luego la calificación de los calificadores de este examen y trabaja con aquellos que se han alejado de la aplicación constante del baremo de calificación. Otro método es el de celebrar pequeñas sesiones de formación a primera hora de la mañana o inmediatamente después del almuerzo. Los calificadores corrigen un conjunto común de exámenes

de control, igual que hicieron durante su formación. Aquellos que se han alejado del estándar preestablecido (acuerdo total: más o menos un punto) participarán en una sesión de repaso o volverán a ser controlados antes de permitirles seguir calificando.

Otra consideración de constancia en la evaluación a gran escala se refiere al control de prejuicios en las decisiones de calificadores. Habrá que asegurarse de que los calificadores que trabajan juntos no formen subgrupos según afinidades de acuerdo ignorando al resto de los calificadores. Para evitar esto, se deberá romper los grupos de calificadores a intervalos periódicos y volver a puntuar exámenes/trabajos ya calificados por otros calificadores asignados a otras mesas o localidades.

La logística de la organización

Aunque la preocupación más importante del proceso de evaluación es la de lograr la constancia, llevar a cabo una sesión de calificación implica varias cuestiones logísticas y técnicas. Buscar la hora más apropiada es una de las cuestiones fundamentales en la planificación de una sesión de calificación. Puesto que la tendencia natural es la de sentirse cansado conforme el día avanza, sería conveniente programar esta sesión a una hora temprana y así evitar las últimas horas de la tarde. Un fácil acceso a una fotocopiadora solucionaría cualquier falta inesperada de material o incluso permitiría copiar aquellos exámenes que deben discutirse durante la sesión. Además, la corrección es una actividad muy intensa; se deberá programar frecuentes descansos y refrigerios (mucha fruta y carbohidratos, y pocos azúcares). El lugar donde se lleve a cabo la sesión debe ser tranquilo y cómodo, con mucho espacio para que los calificadores puedan acomodar todo el trabajo que se va a evaluar. La pesadilla de los calificadores es trabajar en un gimnasio con mesas y sillas plegables a las 3:30 de una tarde calurosa de Mayo y con la banda del colegio ensayando en el patio.

Otra cuestión es la de controlar la distribución de los exámenes o trabajos. En las evaluaciones a gran escala, cada mesa de calificadores debe tener su propio coordinador cuya única preocupación es la de dirigir la distribución de los exámenes y controlar y vigilar la constancia de los calificadores. Nuestra experiencia indica que las pilas de trabajos que llevan alrededor de una hora para su corrección resultan más fáciles para los calificadores que trabajos individuales. El número de trabajos en cada montón varía según la naturaleza de la tarea y la complejidad del sistema de calificación. En evaluaciones de expresión escrita, por ejemplo, podemos incluir entre 15–25 exámenes, mientras que en un conjunto de carpetas de trabajo se debe incluir tan sólo 4–6. Sin tener en cuenta cómo se va a agrupar el trabajo, se deberá asignar aleatoriamente los trabajos individuales a los montones para luego asignar éstos, también de forma aleatoria, a los calificadores para así evitar que se produzcan efectos de calificación sistemática. Para evaluaciones formales, se deberá asignar números de identificación tanto a calificadores como a alumnos, y de esta forma proteger la intimidad y evitar prejuicios.

97

Habrá también que decidir si se debe mezclar cursos o temas dentro de una misma sesión de calificación. Generalmente no se hace, a menos que el objetivo de la evaluación sea el de comparar alumnos de distintos cursos con el mismo baremo de calificación. En las evaluaciones a gran escala, los distintos temas se asignan bien a diferentes grupos de calificadores o se califican por separado con una sesión previa para refrescar la memoria del calificador antes de cada cambio de tema.

Otro asunto que puede dar problemas más adelante si no se controla con cuidado, es asegurar que los calificadores anoten la información necesaria de forma correcta. ¿Se han rellenado todas las casillas con los números de identificación y las calificaciones? ¿Se han anotado las calificaciones para todos los exámenes corregidos? ¿Tienen todos los alumnos sus calificaciones? La lista es enorme. Se debe intentar anticipar los posibles problemas y crear estrategias sea para prevenirlos o para solucionarlos.

Asegurar una calidad técnica

Aconsejar sobre todas las decisiones técnicas que se deberá tomar para asegurar la exactitud y equidad de la calificación no entra dentro de los objetivos de este libro y, de cualquier forma, correspondería al campo de la psicometría. Si se está evaluando con el fin de tomar una decisión trascendental, y especialmente si esa decisión puede ser motivo de demanda, aparecer en la primera plana de un periódico local, o pasar por un comité de educación, quizá se desee recurrir a la ayuda de un asesor técnico que estructure el proceso de calificación y ayude a documentar la fiabilidad de las notas dadas a los alumnos. Algunas de las cuestiones que habrá que abordar son las siguientes:

¿Cuántos calificadores hacen falta? Por supuesto, esto depende del número de trabajos que se van a corregir, de cuántas correcciones reciba cada trabajo, del tiempo que se tarde en corregir cada trabajo y del número de días de que se dispongan para la calificación. En general la calificación integral de redacciones de una a dos páginas es rápida, a veces incluso a examen por minuto. Una evaluación compleja analítica de trabajos más largos puede llevar de cuatro a cinco minutos por examen. Las carpetas de trabajos pueden incluso llevar más tiempo. En cuanto al número de días, nuestra experiencia indica que los calificadores pueden estar muy cansados después de cuatro o cinco días.

¿Cuántas correcciones por examen? Una corrección eficaz y una vigilancia y control cuidadosa del proceso de calificación pueden disminuir la necesidad de llevar a cabo una corrección múltiple de la misma dimensión del trabajo del alumno. Los calificadores múltiples son necesarios para cada examen cuando tienen poca experiencia o cuando existen pocas pruebas de que no estén empleando los mismos criterios y estándares a la hora de tomar decisiones. La necesidad de correcciones múltiples depende de la finalidad de la evaluación. Entre más graves sean las consecuencias, más importante será asegurar la constancia. Nuestra experiencia indica que no se necesitan más de dos calificadores por trabajo; las calificaciones se pueden sumar o se puede sacar la nota media para llegar a la nota final. Se puede recurrir a una tercera opinión para los casos difíciles, como por ejemplo para ese examen de pesadilla que saca tanto la calificación más baja como la más alta.

En algunas situaciones, una corrección es suficiente para la mayoría de los exámenes. Consideremos una situación en la que la selección, asignación un curso específico u otra decisión crítica se va a tomar basándose en algún estándar o calificación preespecificada. Si los exámenes utilizados en la formación y los del control demuestran que los calificadores son constantes, los únicos exámenes que requerirán dos o más correcciones serán aquellos exámenes dudosos que tengan una calificación rozando el aprobado. Puesto que la corrección es un proceso caro, habrá que lograr el equilibrio entre cuestiones como la fiabilidad y aquellas de coste y eficacia.

¿Cómo se califican los exámenes con fines evaluativos? Si las calificaciones de los alumnos se van a utilizar para la evaluación del programa en lugar de para una evaluación individual, una estimación fiable de la calificación de un determinado alumno es menos crítica que la calificación media de la tarea. La mayoría de los trabajos sólo se pueden leer una vez y las pruebas de fiabilidad sólo pueden obtenerse en una única muestra de trabajo (quizás un 20 por ciento), que la corrigen dos o más calificadores. Si se está utilizando muestras de alumnos para evaluar un programa y no hay que proporcionar calificaciones individuales a los maestros, es más eficaz calificar una muestra seleccionada aleatoriamente. El asesor técnico puede aconsejar sobre el tamaño de la muestra y la manera apropiada de seleccionarla.

Proporcionar pruebas de fiabilidad

Para evaluaciones trascendentales, se deberá documentar oficialmente la constancia y fiabilidad del proceso de calificación. Es conveniente aprovechar los servicios de un experto técnico antes de empezar a calificar para así asegurarse de que se tiene un esquema de calificación apropiado, de que se está recopilando pruebas apropiadas y de que los datos utilizados se han formateado de una manera apropiada para facilitar su análisis.

A continuación presentamos algunas de las fuentes de pruebas relevantes:

- **Los resultados del control de calificadores después de su formación.** Se deberá informar sobre el nivel de acuerdo exigido. ¿Cuál fue la proporción de calificadores aprobados en un primer momento? ¿Cuál fue el nivel medio de acuerdo entre aquellos que aprobaron?
- **Los resultados del control de constancia durante la corrección.** Se deberá informar sobre el nivel de acuerdo exigido. ¿Cuántos controles hubo y cuándo se llevaron a cabo? ¿Cuál fue la proporción de los calificadores que aprobaron sin tener que repetir el proceso? ¿Cuál fue el nivel medio de acuerdo en los controles?
- **Los resultados de fiabilidad entre calificadores en la evaluación de trabajos de alumnos llevada a cabo por más de un calificador.** El acuerdo expresado a manera de porcentaje entre los calificadores y los coeficientes de generalizabilidad son dos técnicas usadas frecuentemente. Cada uno de ellos se calcula por separado para cada baremo empleado. Como guía se

necesita una corrección doble al menos en un 20 por ciento de las muestras de alumnos para poder recopilar suficientes evidencias y, si están implicados más de dos calificadores, haría falta consultar a un estadístico que ayude a crear un diseño equilibrado que especifique cuáles son los trabajos que debe corregir cada calificador.

¿Qué nivel de acuerdo o fiabilidad es el deseable? Por supuesto la respuesta es: depende de las decisiones que se tomen. Cuanto más críticas o restrictivas sean las consecuencias, más fiable ha de ser la calificación. En general, los coeficientes de fiabilidad de .70 en adelante se consideran adecuados. Los coeficientes de .90 en adelante son frecuentes en exámenes tipo test estandarizados y evaluaciones directas de la expresión escrita a gran escala.

- **La constancia de los calificadores con el paso del tiempo.** Cuando se quiere asegurar que el baremo de calificación sea constante año tras año—por ejemplo, cuando se utilizan resultados en evaluaciones estatales para seguir las tendencias con el paso del tiempo—se deberá incluir junto con la evaluación de este año una muestra adecuada de trabajos de alumnos de la evaluación del año anterior. El acuerdo entre las calificaciones dadas se puede comprobar posteriormente y, si es necesario, se puede hacer ajustes estadísticos con las diferencias.
- **La constancia entre los calificadores de diferentes lugares o entre diferentes grupos de calificadores.** Al igual que en el control de la constancia con el paso del tiempo, si el trabajo del alumno va a ser calificado en lugares distintos o por grupos diferentes de calificadores, habrá que llevar un control de constancia de estos grupos. Por ejemplo, un estado puede convocar cuatro talleres regionales para calificar sus evaluaciones prácticas de ciencias, o una evaluación a nivel distrito puede exigir que cada escuela califique los trabajos de sus alumnos. Una manera de comprobar la constancia podría hacerse comparando el trabajo calificado por cada grupo con un conjunto común de trabajos. En el lugar número uno, por ejemplo, los calificadores evaluarían los trabajos asignados específicamente a ese lugar más los del conjunto común; los calificadores del lugar número dos evaluarían los trabajos asignados a ese lugar más los del conjunto común, etcétera. Las calificaciones para el conjunto común serán posteriormente comprobadas para averiguar su constancia.
- **La constancia entre los calificadores.** Este es el nivel de constancia que mantiene un calificador a lo largo del tiempo. Esto se puede comprobar pidiendo a los calificadores que califiquen un mismo trabajo más de una vez en distintos momentos del proceso de calificación.

Comprobar la fiabilidad del proceso de corrección

Como resumen de muchas de las cuestiones tratadas en este capítulo, se puede utilizar la siguiente lista de control para averiguar si los procedimientos de calificación son buenos y fiables. Se cuenta con:

- [ ] guía de calificación documentada y probada
- [ ] criterios claros y concretos
- [ ] ejemplos anotados de todos los valores de calificación
- [ ] oportunidades de practicar con retroalimentación para los calificadores
- [ ] calificadores múltiples que demuestran un acuerdo previo al proceso de calificación
- [ ] controles periódicos de fiabilidad durante todo el proceso
- [ ] volver a capacitar al calificador cuando sea necesario
- [ ] previsiones para la recopilación de datos apropiados de fiabilidad

Referencias bibliográficas

Baker, E.L., P.R. Aschbacher, D. Niemi, y E. Sato. (1992). *CRESST Performance Assesment Models: Assessing Content Area Explanations*. Los Angeles: University of California, Center for Research on Evalution, Standars, and Student Testing.

Dreidrich, P.B. (1963). "The Measurement of Skill in Writing." *School Review* 54: 584-592.

Finch, F. (1991). *Educational Performance Assessment*. Chicago: Riverside Publishing Company.

Quellmalz, E., y J. Burry. (1983). "Analytic Scales for Assessing Students' Expository and Narrative Writing Skills." (CSE Resource Paper No. 5). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Students Testing.

7

Utilización de la evaluación alternativa para la toma de decisiones

A lo largo de este libro hemos examinado varias cuestiones importantes en el diseño de evaluaciones alternativas de calidad: ¿Qué es la evaluación alternativa? ¿Cómo podemos identificar tareas apropiadas de evaluación? ¿Qué debieran incluir los criterios? ¿En qué consisten los buenos procedimientos de calificación? Ahora volvemos nuestra atención a la razón que nos condujo en un principio al diseño de evaluaciones alternativas: la de tomar decisiones apropiadas sobre alumnos y programas.

Este es un punto de importancia clave: la evaluación no es una finalidad en sí misma. Más bien, la evaluación proporciona información para poder tomar decisiones sobre lo que los alumnos han aprendido, qué notas se merecen, si los alumnos deben o no pasar al siguiente curso, a qué grupos se les debe asignar, qué ayuda necesitan, qué áreas de la didáctica de la clase necesitan renovarse, si el currículo del centro necesita reforzarse, etcétera. Una buena evaluación nos permite caracterizar con exactitud el funcionamiento y rendimiento de los alumnos para poder tomar decisiones apropiadas que mejoren la educación.

¿Contribuye la utilización de los resultados de una evaluación a tomar decisiones acertadas? Esta es la clave de cómo valoramos la calidad de una evaluación. Los coordinadores de educación y el público en general tienen mucha fe en las pruebas estandarizadas, y en su eficacia en ayudarnos a sacar conclusiones acertadas sobre alumnos y centros. Desgraciadamente, algunos creen que esta fe no tiene base. Al convertirnos en consumidores de evaluaciones más sofisticados, nos hemos cuestionado más sobre lo que realmente nos dicen estas pruebas. ¿Las calificaciones de la Prueba de Aptitud Escolar (Scholastic Aptitude Test) identifican realmente a los alumnos que van a tener éxito en la universidad? Si no fuera así, ¿cuánto peso se les debe dar en las decisiones sobre el ingreso a la universidad? ¿Proporcionan las evaluaciones estatales el tipo de información que los centros necesitan para mejorar sus programas? ¿Ayudan a los coordinadores de educación y al público en general a averiguar si los alumnos aprenden lo que necesitan saber y ser capaces de hacer? ¿Los exámenes tipo test permiten a los alumnos demostrar su plena comprensión de una asignatura? Si no fuera así, ¿hasta qué punto debiéramos depender de ellos cuando se toma decisiones sobre alumnos y programas?

La insatisfacción con los exámenes tradicionales ha alentado a los maestros y al estado entero a encontrar formas alternativas de evaluación. Sin embargo, los formatos alternativos por sí solos no pueden garantizar una buena evaluación. Debemos aplicar a las evaluaciones alternativas el mismo escrutinio que nos permitió ver tanto las limitaciones como los puntos fuertes de los exámenes tradicionales. Tenemos que asegurarnos de que las evaluaciones que pretendemos utilizar vayan a ayudar y no a perjudicar a los alumnos, programas y centros.

Este capítulo destaca los temas que deben tomarse en cuenta cuando se utilizan las evaluaciones, sean alternativas o no. Empezamos con un análisis a dos conceptos claves para evaluar la calidad de cualquier evaluación: la validez y la fiabilidad. Más adelante examinamos tres importantes cuestiones que guían la utilización apropiada de la información que arroja la evaluación:

1. ¿Cómo influye el contexto de la decisión que se ha de tomar y su utilización intencionada a las cuestiones de calidad del programa de evaluación?
2. ¿Cómo podemos asegurarnos de que una evaluación nos proporciona la información apropiada para la toma de decisiones?
3. ¿Cómo podemos utilizar los resultados de la evaluación para mejorar la enseñanza?

Vean cómo nos dirigimos a cuestiones de calidad evaluadora antes de proporcionar ejemplos concretos de cómo utilizar los resultados de una evaluación. Lo hacemos así para enfatizar que la calidad de la evaluación siempre es importante y que hay que tenerla en cuenta antes de utilizar los resultados. Si una evaluación no proporciona buena información para tomar decisiones, su utilización puede ser perjudicial.

Antes de aventurarnos más, recordamos a los lectores que por razones de simplificación a lo largo de este libro hemos examinado cuestiones desde la perspectiva de una única evaluación. Sin duda los lectores sabrán perfectamente

que ninguna evaluación o examen individual constituye una buena estrategia de evaluación. Todas las evaluaciones, incluso las mejores, son imperfectas y falibles. Las evaluaciones alternativas, como todas las evaluaciones, deben utilizarse junto con otras fuentes de información para constituir un programa de evaluación sistemático y equilibrado. Mientras leemos acerca de los factores que influyen en la utilización de exámenes, hay que recordar que las mismas cuestiones que corresponden a una evaluación individual se pueden aplicar a todo un conjunto de evaluaciones o a un sistema completo de evaluación.

Cuestiones que aseguran la calidad, la validez y la fiabilidad

¿Proporciona una evaluación la información precisa para la toma de decisiones? ¿Permiten sus resultados sacar conclusiones precisas y justas sobre el rendimiento del alumno? ¿Nos conduce la utilización de los resultados a unas buenas decisiones? Estas son las cuestiones centrales cuando juzgamos la calidad de una evaluación. Si queremos obtener respuestas afirmativas a estas preguntas, nuestras evaluaciones han de ser tanto fiables como válidas—términos que utiliza la comunidad de medición para estas mismas cuestiones.

La fiabilidad: La estabilidad del rendimiento

Anteriormente introdujimos el concepto de fiabilidad en relación con la constancia de juicios humanos. Hemos visto que existen varias formas para asegurar niveles aceptables de acuerdo sobre el rendimiento escolar entre calificadores. Sin embargo, la fiabilidad en su sentido más amplio corresponde a si los resultados de unos exámenes mantienen su significado (permanecen constantes) a pesar de la existencia de cambios superficiales en la situación evaluadora—de un día a otro, sin tener en cuenta la persona que evalúa el rendimiento o el día o la hora en la que se lleva a cabo la calificación. Si María escribe un comentario sobre Tristam Shandy hoy, mañana o el martes que viene, se espera que su rendimiento sea esencialmente el mismo en las tres ocasiones. Si su maestro lee su trabajo esta noche, mañana o el martes que viene, se espera que le dé la misma calificación o que saque las mismas conclusiones sobre el desarrollo de sus destrezas y sus puntos fuertes o débiles. Si Byron es capaz de crear dos métodos para resolver un problema de matemáticas hoy, se espera que sea capaz de hacer un análisis parecido para un problema similar el viernes o la semana que viene. Sin esta constancia, no podemos decir con seguridad lo que un alumno es capaz de hacer. Una calificación no fiable es inútil porque no nos diría nada significativo o generalizable sobre el rendimiento escolar. Por esta razón, debemos asegurarnos de que nuestros resultados sean fiables antes de preocuparnos por la validez, tema más relacionado con la utilización de exámenes. De hecho, la mayoría de nosotros hemos aprendido en algún momento la máxima "para ser válida, una calificación ha de ser fiable". Cuando se nos pide recordar este tema, muchos de nosotros no estamos seguros si es que la fiabilidad

precede a la validez o viceversa. Quizá la manera más fácil de recordar este orden sería acordándose de que si una calificación va a ser valiosa (validez) para la toma de decisiones ha de ser factible (fiable).

La validez: La exactitud de las conclusiones basadas en exámenes

Los especialistas de medición saben que aunque la fiabilidad es necesaria, no es condición suficiente para la validez—en otras palabras, si el resultado de un examen da lugar a conclusiones exactas sobre el rendimiento de un alumno y es, por consiguiente, una buena base para la toma de decisiones. El resultado de un examen podría ser perfectamente fiable pero no relevante para la decisión que se pretende tomar. Por presentar un ejemplo extremo, una prueba de mecanografía o de procesador de textos puede proporcionar información muy fiable (repetible y constante) para valorar las destrezas y la velocidad cuando se escribe a máquina, pero estos resultados son inútiles en la toma de decisiones sobre la capacidad del alumno en la expresión escrita. Igualmente, una prueba de multiplicar puede ofrecernos información fiable acerca de las destrezas de computación de los alumnos, pero sería poco útil para determinar si saben resolver problemas.

La determinación de la validez de una evaluación depende de la forma en que se pretende utilizarla. A lo largo de este libro hemos utilizado un tanto libremente la palabra "validez", como si fuera una calidad o característica de un examen en particular. En realidad, las propias evaluaciones no son ni válidas ni inválidas; su validez depende de los fines para los que las utilizamos. Evaluamos la validez de un examen determinando si una conclusión basada en la calificación dada al examen es apropiada para un objetivo en particular o no. Por ejemplo, si queremos utilizar los resultados de un examen para identificar a aquellos alumnos que dominan las ecuaciones lineales nos preguntaremos, ¿Identifican las calificaciones recibidas a todos los alumnos que dominan las ecuaciones lineales? o ¿Los alumnos que han sido identificados como alumnos que necesitan ayuda la necesitan realmente? Para ser más precisos, cuando hablamos de la validez del examen en la identificación de alumnos que dominan ecuaciones lineales, en realidad estamos haciendo referencia a la evidencia que tenemos que nos indica que nuestras conclusiones basadas en la calificación son correctas, que los alumnos que obtienen un aprobado, o más, realmente dominan el contenido. Pocas son las razones que tenemos para utilizar resultados y puede que haya poca seguridad al hacerlo hasta que tengamos evidencias que lo corroboren, como por ejemplo el rendimiento del alumno en trabajos posteriores, el rendimiento en evaluaciones parecidas, la observación de maestros y otras decisiones de maestros que apoyen nuestras conclusiones basadas en calificaciones.

Puesto que sería bastante molesto repetir esta definición tan precisa, a continuación utilizaremos "validez" para "la evidencia que respalda las inferencias basadas en la calificación". Conforme se lee, se debe tener en mente la definición más exacta.

Hay que recordar también que las evaluaciones pueden ser válidas para algunos objetivos pero poco apropiadas para otros. Por ejemplo, un examen diagnóstico de las destrezas básicas proporciona comparaciones de gran utilidad con una muestra nacional, pero puede resultar relativamente inútil en la identificación del dominio de objetivos curriculares locales. Los resultados de un examen final pueden ser válidos para decidir si un alumno debiera recibir una "A" o una "B" en una clase, pero pueden no serlo en la identificación de aquellos alumnos que sacarían más provecho de una enseñanza acelerada o de aquellos alumnos especiales que podrían participar en el nuevo programa de alumnos superdotados. Lo que debemos concluir de ello es que si un examen presume de tener múltiples usos, hay que acompañarlo de evidencias que apoyen cada uno de los usos. ¿Qué tipo de evidencias son? El siguiente apartado proporciona lo que hay que tener en cuenta cuando se decide qué tipo de evidencias formales se querrá considerar al utilizar evaluaciones para tomar decisiones sobre alumnos, clases o centros.

¿Cómo afectan el contexto de las decisiones y su utilización intencionada a las cuestiones de calidad?

Conocer la finalidad de la evaluación

Las evaluaciones se diseñan con el fin de proporcionar información para la toma de decisiones acerca de alumnos, clases, centros, distritos, estados y objetivos educativos nacionales. ¿Cuál es la finalidad de su evaluación? ¿A qué público van dirigidos los resultados? ¿Qué otra información utilizará estos públicos para sacar conclusiones o para tomar decisiones? Las respuestas a estas preguntas tienen implicaciones serias en lo que se refiere al contenido que se debe incluir en una evaluación, cómo debe realizarse y cuánta atención se debe prestar a asegurar su calidad.

Importancia de las consecuencias

Es evidente que algunas decisiones sobre alumnos y centros conllevan consecuencias más graves que otras. Los exámenes trascendentales conllevan consecuencias importantes. Las evaluaciones no trascendentales tienen un impacto menor en los individuos. Estas incluyen evaluaciones que se utilizan para la monitorización del progreso, programación la enseñanza, e incluso para calificar cursos determinados (si se va a utilizar una variedad de notas y otras evidencias para sacar la nota final). Cuanto más trascendental es la evaluación, mayor es la necesidad de documentar su calidad—su validez y fiabilidad.

Recopilación de evidencias que corroboran la toma de decisiones

Incluso en situaciones no trascendentales, se puede cometer errores que causen mucho daño. La acumulación de exámenes por tema y otras evaluaciones llevadas a cabo en el aula envían importantes mensajes a los alumnos y a los padres, y pueden tener un impacto significativo en ellos. De igual manera, las valoraciones informales de la calidad escolar basadas en resultados de evaluaciones pueden afectar a la moral así como a las actividades del cuerpo docente con el paso del tiempo. Por consiguiente, la validez necesita de nuestra atención sin tener en cuenta si el contexto de la evaluación es trascendental o no.

Si identificamos puntualmente si la evaluación es o no trascendental nos ayudará a determinar cuántas evidencias necesitamos para documentar la calidad de la evaluación. ¿Cuáles son las consecuencias del rendimiento en el examen? ¿Se van a utilizar los resultados de la evaluación junto con otra gran cantidad de información corroborante para tomar decisiones acerca de los alumnos? ¿Será ésta prácticamente la única base para una decisión? ¿Si una decisión basada en la calificación es incorrecta, se puede corregir? ¿Puede estar sujeta a una demanda? Si una evaluación conlleva consecuencias importantes como lo hacen casi todas aquellas utilizadas para la responsabilidad adjudicada o rendir cuentas (accountability), la adjudicación de plazas, o financiamiento, es imprescindible tener evidencias formales de validez para los objetivos establecidos.

Pruebas de validez: ¿Cómo se sabe si una evaluación nos proporciona buena información?

A lo largo de este libro se ha tratado asuntos relacionados con la validez de la evaluación, por lo que algunos de los temas que vamos a destacar a continuación resultarán familiares. Evidentemente, la calidad o validez de una evaluación con un fin determinado depende de varios puntos y requiere que se tome en cuenta una serie de evidencias. Aquellos que están interesados en obtener más detalles técnicos y en técnicas para recopilar evidencias corroborantes quizás les interese consultar Standards for Educational and Psychological Tests (1985). Los criterios descritos en este trabajo sirven como piedra de toque para la calidad de exámenes cada vez que se cuestione una evaluación en un juicio. Seguirlos de cerca nos proporciona la seguridad de que se pueda defender en caso de litigio cualquier evaluación que se utilice.

¿Pueden utilizarse las calificaciones para describir lo que los alumnos han aprendido?

Unos de los primeros fines de la evaluación es el de averiguar lo que los alumnos saben o han aprendido con respecto a determinados objetivos didácticos. La validez para tal propósito requiere que exista una buena relación entre estos objetivos y el contenido de la evaluación. Las siguientes preguntas ayudarán a decidir si existe tal relación:

- ¿Va el examen acompañado de una clara definición de los objetivos de la evaluación de manera que se pueda valorar la relación entre las destrezas y los conocimientos que se pretende evaluar, y aquellos enfatizados en la clase o en el centro?
- ¿Refleja el contenido de la evaluación el contenido más importante y completo del currículo? ¿Existe una buena relación entre la descripción de la tarea y las prioridades didácticas?
- ¿Las tareas de evaluación requieren el tipo de conocimientos, razonamiento, resolución de problemas y destrezas de procesos que se incluyen en la enseñanza?
- ¿Explota la evaluación las destrezas de razonamiento complejo? ¿Cuáles?
- ¿Incluye la evaluación criterios de calificación? Si es así, ¿corresponden estos criterios a los objetivos didácticos, las teorías de aprendizaje actuales y las prioridades curriculares?
- ¿Incluyen los criterios estándares que valoren el nivel de rendimiento del alumno? Si es así, ¿cómo se establecieron estos criterios?
- ¿Es la tarea, desde el punto de vista del desarrollo del alumno, apropiada? ¿Refleja los procesos y objetivos apropiados para los alumnos a los que está dirigida?
- ¿Han tenido los alumnos la suficiente oportunidad para aprender lo que se incluye en la evaluación?

Si las respuestas a estas preguntas son afirmativas, tendremos entonces evidencias de que los resultados de la evaluación nos conducirá a conclusiones fiables acerca de lo bien que los alumnos han logrado los objetivos educativos, al igual que lo eficaz que ha sido la enseñanza.

Si deseamos más evidencias de la validez de un examen en estas dimensiones, podemos pedir a un colega que nos revise la evaluación y que se plantee las mismas preguntas, o bien preguntas menos específicas como pueden ser las siguientes:

1. ¿Qué es lo que en su opinión mide esta evaluación?
2. ¿Qué es lo que esta evaluación, con respecto a mis objetivos, me va a decir sobre mis alumnos? ¿Sobre los niveles de rendimiento de nuestro centro educativo? ¿Sobre las metas importantes de los alumnos? ¿Sobre los puntos fuertes y débiles de los alumnos?

3. ¿Es este tipo de evaluación la que usted hubiera pensado para evaluar sus objetivos?
4. ¿Cómo sería una respuesta típica a esta evaluación?

Se obtendrán evidencias aún más formales si se convoca a un equipo de expertos en la materia y se le pide que valore la evaluación según los criterios de correspondencia curricular. Cuando se trata de un examen trascendental, como por ejemplo un examen a nivel estatal, hay que obtener este tipo de evidencias.

Al analizar la validez de la evaluación en estas áreas, hay que estar conscientes de los límites de la validez aparente. Aunque pueda parecer que la tarea evalúe los objetivos deseados, no podemos estar completamente seguros de lo que estamos midiendo hasta que veamos las propias respuestas de los alumnos. ¿Cuáles son los conocimientos y destrezas que los alumnos emplean en esta evaluación? La única manera de saber si la evaluación realmente evalúa los objetivos previstos por medio de la recopilación de datos que corroboren la nota dada. Se puede obtener estos datos por medio de la observación, el análisis cuidadoso del rendimiento del alumno, o pidiendo a los alumnos que nos informen sobre las destrezas y conocimientos que utilizaron al hacer la tarea de evaluación. Por ejemplo, si la tarea se ha diseñado para valorar la habilidad del alumno para relacionar la personalidad de Hamlet con otros personajes históricos, no se podrá estar seguro si las respuestas reflejan un razonamiento crítico y la utilización de conceptos en nuevos contextos. Para poder averiguar si la evaluación produce resultados válidos, es necesario asegurarse de que los alumnos no hayan preparado o memorizado respuestas, no hayan utilizado determinadas obras críticas publicadas sobre Hamlet, o que con anterioridad hayan respondido a esta pregunta.

Una vez que hayamos determinado que la evaluación refleja los objetivos previstos, podemos pasar a la importante cuestión sobre hasta qué punto refleja la calificación obtenida el logro de un alumno.

¿Son las calificaciones obtenidas válidas como para generalizar sobre un alumno?

Una cuestión importante al determinar la validez de las evaluaciones de rendimiento de cualquier tipo es saber si se puede generalizar sobre el rendimiento de un alumno a partir de una tarea determinada. Después de todo, enseñamos con el fin de transferir información. Queremos que nuestros alumnos tengan conocimientos y destrezas duraderos. Por ello, esperamos y muchas veces damos por sentado, que el rendimiento del alumno en nuestras tareas evaluativas puede generalizarse a un dominio mayor y que los resultados de una evaluación representan la forma en la que el alumno se va a desempeñar en un conjunto más amplio de tareas. Después de todo, cuando damos un examen práctico de ciencias naturales a los alumnos en el que utilizan gusanos de seda, probablemente no nos importa tanto si los alumnos son capaces de llevar a cabo este experimento en concreto con gusanos de seda como sus destrezas en utilizar el método científico.

La cuestión de transferencia y generalizabilidad parece ser un tema problemático en la evaluación alternativa, donde el tiempo disponible restringe el número de tareas que pueden realizar los alumnos. ¿Qué tareas, destrezas, contenido y rendimientos hay que incluir en una evaluación para asegurarse de que se pueden aplicar a un dominio superior? ¿Cuántas muestras de rendimiento del alumno necesitamos antes de poder hacer estas generalizaciones? No lo sabemos con precisión, pero la respuesta, desgraciadamente, no es sólo una.

Por ejemplo, Herman (1991) examinó la investigación sobre la evaluación de la expresión escrita y descubrió que la destreza de escribir no se puede generalizar a los distintos géneros. Para ser más específicos, aquellos alumnos que escriben buenas redacciones persuasivas no tienen por qué saber escribir buenos cuentos o buenos comentarios literarios. Además, incluso dentro del mismo género, el rendimiento de un alumno puede variar bastante según el tema o estímulo. Estos descubrimientos sugieren que a pesar de la validez intuitiva de las tareas y de hasta qué punto éstas ocupan de manera significativas a los alumnos, las evaluaciones alternativas no conducen necesariamente a inferencias más validas sobre dominios superiores de rendimiento. En otras palabras, parece que hay un equilibrio entre la profundidad y la amplitud de la información que proporciona tales evaluaciones.

¿Cómo sabemos si los resultados obtenidos en la evaluación de un alumno representan un dominio significativo superior de rendimiento? Recopilamos evidencias sobre la generalizabilidad averiguando la constancia del rendimiento del alumno en varias tareas que están diseñadas para evaluar los mismos conocimientos, destrezas y disposiciones. Técnicamente hablando, podemos llevar a cabo análisis estadísticos especiales que cualifiquen la relación entre el rendimiento de una tarea y otra, utilizando luego las reglas de decisiones para determinados controles estadísticos y así decidir si deberíamos confiar en los resultados. Aunque este tipo de análisis no cabe dentro de este estudio, hay que estar conscientes de que en situaciones trascendentales en las que se utilicen exámenes formales se requerirán pruebas estadísticas. Se debe presentar datos formales para poder contestar la pregunta: "¿Basándome sólo en esta tarea, hasta qué punto ès mi decisión sobre un alumno acertada?" O, aún más útil sería la pregunta "¿Cuántas tareas parecidas a éstas tiene que realizar un alumno para yo poder tomar una decisión con un mínimo de exactitud?"

Al reconocer que no es práctico llevar a cabo análisis estadísticos complejos para la mayoría de las evaluaciones efectuadas en el aula, todavía podemos mejorar la validez de nuestras conclusiones sobre los alumnos utilizando todas las observaciones o muestras de trabajos posibles antes de hacer generalizaciones o sacar conclusiones sobre la capacidad de rendimiento de un alumno.

¿Se pueden utilizar las calificaciones para diagnosticar los puntos fuertes y débiles de los alumnos?

¿Se pueden utilizar las calificaciones para diagnosticar los puntos fuertes y débiles del currículo? Otra cuestión sobre la validez que es fundamental en la utilización de evaluaciones en el aula y en el centro educativo es su utilidad diagnóstica. ¿Nos dicen los resultados algo significativo sobre porqué los alumnos han rendido de tal manera?

Si deseamos utilizar las calificaciones para diagnosticar los puntos fuertes y débiles de los alumnos, las tareas y los criterios de calificación deben basarse en una teoría de aprendizaje sobre la adquisición de destrezas o conocimientos con credibilidad. Veamos lo que pasa cuando una supuesta calificación "diagnóstica" no concuerda con la teoría en cuestión. Recientemente, si la expresión escrita de un alumno se consideraba insuficiente, los maestros se centraban en la enseñanza de las destrezas necesarias como son la gramática, la técnica y la estructura de párrafos. La investigación sobre el proceso de escribir desacredita este método de destrezas aisladas al igual que el valor diagnóstico de contar los errores gramaticales y técnicos como indicadores de calidad de escritura (Braddock et al. 1963, Elley et al. 1976). Podemos citar un ejemplo análogo en el área de matemáticas. Si bien es verdad que la automaticidad de calcular ayuda a los alumnos con las matemáticas, puede ser que el dominio de fracciones, decimales y divisiones no ayude al rendimiento del alumno con el álgebra. En resumen, los tests diagnósticos preálgebra que se exige a la mayoría de los alumnos de octavo grado de este país están basados en teorías deficientes sobre la agilidad algebraica. Estos ejemplos ilustran el gran reto que existe en la creación de evaluaciones diagnósticas al igual que ilustran la cautela que debemos tomar al buscar información diagnóstica en nuestras propias evaluaciones.

En capítulos anteriores destacamos la necesidad de coordinar las descripciones de tareas y criterios con las teorías actuales de currículo y aprendizaje. Esta base teórica también proporciona una relación entre los resultados deseados y los requisitos necesarios. Una evaluación válida desde el punto de vista diagnóstico es prueba de un cuerpo de investigación que respalda la unión entre determinadas calificaciones diagnósticas y la teoría subyacente.

¿Es objetiva la calificación dada?

Otra cuestión fundamental sobre la validez de la evaluación del aula y del centro educativo es la de ser justo y objetivo. La teoría reciente de aprendizaje cognitivo destaca la importancia de los conocimientos previos cuando se resuelve problemas. Es evidente que alumnos de distintos entornos socioeconómicos, culturales y lingüísticos pueden tener distintos tipos de conocimientos y experiencia previos. ¿Tienen los alumnos suficientes conocimientos previos para tener éxito en la tarea de evaluación? ¿El contenido o contexto de la evaluación da injustamente ventaja o desventaja a niños de distintos grupos culturales o lingüísticos? ¿Es igual de

significativa y motivadora para alumnos de distintos entornos culturales? ¿Contiene la evaluación material o estereotipos que son culturalmente inapropiados? Las respuestas a preguntas como éstas proporcionan una línea de evidencia acerca de la objetividad o parcialidad de las evaluaciones.

Se puede disminuir los problemas que causan las diferencias de conocimientos previos si estamos seguros de que todos los alumnos en el centro tienen a sus alcance suficientes oportunidades para adquirir los conocimientos y destrezas que se requieren. Los maestros deben asegurarse de que lo que se está midiendo ha sido enseñado y que los alumnos han tenido la oportunidad de aprender el contenido relevante, y de aplicar los procesos deseados. Muchas autoridades educativas opinan que, en situaciones de evaluaciones trascendentales, se debe buscar regularmente evidencias de que se brinden suficientes oportunidades para aprender. Queremos asegurarnos de que todos los alumnos al menos han gozado de las mismas oportunidades para aprender.

Se puede efectuar una variedad de análisis estadísticos que examinen la parcialidad potencial. Esencialmente, estos análisis buscan el rendimiento diferencial entre los subgrupos, teniendo en cuenta varios factores. Aunque pocos son los maestros o practicantes de enseñanza que tienen que llevar a cabo análisis de este tipo, deberían estar conscientes de la existencia de estos análisis, los cuales debieran estar disponibles para su aplicación en los exámenes formales de gran trascendencia.

¿Hay evidencias que corroboren que la evaluación cumple los objetivos previstos?

Como ya debe ser evidente a estas alturas, demostrar que una evaluación es válida para un objetivo requiere la recopilación de datos específicos para demostrar la relación entre los resultados de la evaluación y ese objetivo. En el caso de exámenes formales de gran trascendencia, esto quiere decir que debiera haber estudios específicos para investigar el significado de las calificaciones correspondientes a estos exámenes (Shepard 1991). Por ejemplo, si se utiliza los resultados de una evaluación estatal de carpetas de trabajo de matemáticas para identificar los puntos fuertes y débiles de un centro educativo, el programa de evaluación estatal necesita recopilar evidencias de que se puedan utilizar las calificaciones de esta manera. O, si afirmamos que la carpeta de trabajo, exposición o tesis de alumnos de cursos superiores demuestran un razonamiento crítico y, habilidades de expresión, así como un dominio de cierto contenido, necesitamos evidencias independientes que corroboren esta interpretación de la calificación. Igualmente, si utilizamos los resultados de una evaluación para decidir quién puede matricularse en la asignatura de álgebra, necesitamos evidencias independientes de la relación entre el contenido del examen, la agilidad algebraica y el rendimiento durante el curso.

¿Tiene la evaluación consecuencias positivas para el aprendizaje y la enseñanza?

La actual polémica sobre los exámenes tradicionales estandarizados nos debiera enseñar una importante lección: tenemos que vigilar las consecuencias de una evaluación. Las buenas intenciones no aseguran resultados positivos. La intención de accountability basada en exámenes fue la de ayudar a mejorar a los centros educativos y su nivel de eficacia con los alumnos. Para muchos, una excesiva dependencia en los exámenes tipo test ha dañado el proceso educativo y se ha alejado de la enseñanza y del aprendizaje significativo.

Queremos asegurarnos de que nuestras nuevas evaluaciones ayuden y no perjudiquen a los centros educativos, y a sus miembros. Para los programas de evaluación obligatorios y trascendentales, esto implica una continua atención a los efectos de los programas y a los estudios formales para evaluar sus consecuencias en el currículo, enseñanza, aprendizaje del alumno, entre otras consecuencias intencionadas o no. Para un maestro en el aula, implica una mayor atención a las consecuencias de la evaluación, por ejemplo:

- ¿Qué valores se ven implicados en la evaluación? ¿Fomenta un razonamiento cuidadoso y la precisión en lugar de la impulsividad? ¿Soluciones múltiples en lugar de una única respuesta? ¿Respeta la diversidad?
- ¿Está bien empleado el tiempo que alumnos y maestros dedican a la preparación de esta evaluación?
- ¿Merecen la pena los objetivos? ¿Se mantiene a los alumnos en un nivel alto? ¿Requiere la tarea la utilización compleja, rica y desafiante de la mente de los alumnos?
- ¿Resultan las tareas auténticas y significativas para los alumnos? ¿Pueden ver los alumnos los vínculos con la vida real?

Repetición: Asegurar la fiabilidad y la validez

Una vez más, queremos tener confianza en la calidad de una evaluación antes de utilizarla. El cuadro 7.1 resume algunas de las estrategias que se tratan en este y en anteriores capítulos que contribuyen a tener esta confianza

¿Cómo podemos utilizar los resultados de la evaluación para mejorar la enseñanza?

Aunque el camino hasta llegar aquí ha sido difícil, por fin hemos llegado con evaluaciones de alta calidad, apropiadas a los fines prefijados. ¿Cómo las utilizaremos? La mayoría de las veces utilizaremos los resultados de evaluaciones para contestar dos preguntas fundamentales:

Cuadro 7.1
Asegurar la fiabilidad y la validez en las evaluaciones alternativas

Etapa en el diseño del examen	Estrategias para asegurar inferencias de una calificación válida
Identificación de los objetivos de la evaluación	• Unir las metas con objetivos curriculares importantes relacionados con contenido, destrezas, procesos transferibles o fundamentales • Crear enunciados claros y sin ambigüedades sobre las metas
Creación de descripciones de tareas	• Crear descripciones de tareas totalmente desarrolladas • Comparar descripción de tarea y metas
Selección/diseño de criterios	• Comparar criterios con metas y teoría subyacente de aprendizaje didáctico o curricular • Asegurar que los criterios reflejen metas que se pueden enseñar • Asegurar que los criterios no favorezcan un determinado sexo, origen étnico, entorno lingüístico
Rendimientos/productos/procesos y la calificación	• En el aula: calificar sistemáticamente y revisar el trabajo con regularidad • Calificar a la vez temas y dimensiones parecidos • Uso a gran escala: formar a los calificadores, vigilar y controlar la constancia • Documentar los varios tipos de fiabilidad (entre calificadores, del mismo calificador, según los temas, según el contexto, con el paso del tiempo para los alumnos) • Asegurar niveles mínimos de "fiabilidad" (del tipo adecuado) y un coeficiente de fiabilidad de al menos .70 para la mayoría de las evaluaciones, .90 para exámenes trascendentales
Utilización de evaluaciones alternativas	• Limitar las inferencias basadas en calificaciones al uso para el que fue diseñada la evaluación o para el que se encuentran múltiples fuentes de evidencias que determinan que se puede utilizar la calificación de una manera determinada • Buscar evidencias en el manual de la evaluación, en trabajos de investigación, en los colegas que apoyen las inferencias basadas en calificaciones • Comprobar inferencias basadas en calificaciones con otro tipo de información, con su experiencia previa, otras calificaciones, otro trabajo del alumno, observaciones • No tomar jamás una decisión importante basada sólo en una calificación

- ¿Qué tal vamos?
- ¿Cómo podemos mejorar?

Intentamos contestar estas preguntas a muchos niveles, desde respuestas sobre alumnos individuales a otras sobre el centro educativo, el distrito escolar, el estado o incluso la nación. Por ejemplo, a un nivel individual: ¿Qué tal va Kang en matemáticas? Y según nuestra respuesta, ¿cómo podemos ayudarle a mejorar? ¿Cómo va Clarissa en ciencias? ¿Y de qué nos sirve la respuesta para saber qué asignaturas le van a beneficiar más el próximo año? O a un nivel de clase, ¿qué tal les fue a mis alumnos en las pruebas orales? ¿Qué me dice la respuesta sobre los puntos fuertes y débiles de mi didáctica en esta área? ¿Necesita una parte de la clase o su totalidad clases de recuperación? O a nivel de escuela, ¿qué tal le fue al quinto grado en las distintas pruebas de expresión escrita? ¿Qué sugieren los resultados de este análisis con referencia a los puntos débiles de nuestro currículo y materiales didácticos?

En los siguientes apartados analizamos los métodos básicos para poder contestar cada una de estas conocidas preguntas.

¿Qué tal vamos?

Establecer estándares

En la pregunta "¿qué tal vamos?" están implícitas cuestiones de calidad y estándares. No sólo queremos saber cómo les va a los alumnos, sino lo que aún es más importante, ¿están los alumnos logrando los objetivos previstos? ¿Están haciéndolo bien? ¿Lo están haciendo tan bien como esperábamos? En pocas palabras, "¿lo estamos haciendo bien, o al menos aceptablemente?"

¿Cómo decidimos la respuesta a este tipo de preguntas? Lo ideal, al formular los criterios de calificación, es tomar en cuenta los niveles de rendimiento. Por ejemplo decidimos que un "5" equivalía a un sobresaliente y un "3" a un simple aprobado. Si este fuera el caso, se puede responder a la pregunta de "¿cómo vamos?" consultando los estándares en los criterios de calificación. Si los criterios son descriptivos y no incluyen niveles de rendimiento, es el momento de atribuir calificaciones específicas a estándares de rendimiento. Hay dos tipos básicos de estándares o comparaciones: absolutos y relativos. Los absolutos prevalecen cuando decidimos qué tanto están rindiendo los alumnos al consultar algún criterio de rendimiento satisfactorio. A veces este criterio lo establece oficialmente un centro educativo o distrito; otras veces es un estándar basado en la disciplina. Los maestros de matemáticas están de acuerdo en lo que se debe incluir en las respuestas a ejercicios matemáticos. Los de lengua inglesa coinciden en los estándares de lo que constituye un resumen bien escrito. Los maestros de ciencias sociales saben qué evidencia es aceptable al respaldar una postura política. Utilizamos estos estándares cuando contestamos preguntas como: ¿fue capaz

Leticia de escribir un buen trabajo de investigación? ¿Fue Judd capaz de calcular los costes de abrir un restaurante?

Podemos utilizar también estándares relativos para evaluar hasta qué punto los alumnos están rindiendo bien. Los estándares relativos son siempre aquellos que comparan los rendimientos de nuestros alumnos con otros grupos de alumnos. Comparar alumnos con la norma nacional (por ejemplo, la calificación en el 50 percentil lograda por una muestra nacional de alumnos) es un ejemplo corriente de un estándar relativo. Los maestros con experiencia normalmente comparan a sus alumnos con otros grupos que conocen bien cuando evalúan el rendimiento. Pueden tener una idea bastante buena del rendimiento a nivel de curso y de la conducta típica de alumnos basándose en la de clases anteriores, o comparando con las clases de sus colegas, o incluso con los resultados de datos de evaluaciones estatales y nacionales. Los estándares relativos nos ayudan a contestar preguntas como: ¿ayudaron los nuevos materiales a que los alumnos mejorasen de alguna manera con respecto al año pasado? ¿Se están desarrollando las destrezas de alfabetización de John a un ritmo aceptable comparado con las normas de desarrollo? ¿Los alumnos del currículo interdisciplinario se están desempeñando tan bien o mejor que los del currículo normal? Si vamos a asignar una calificación, utilizamos muchas veces estándares relativos cuando comparamos el rendimiento actual con anteriores niveles de rendimiento de otros alumnos.

Si bien son útiles en ciertas ocasiones, los estándares relativos tienen graves limitaciones. Su valor se ve limitado por la semejanza entre los grupos que se están comparando. Por ejemplo, sería injusto e inapropiado comparar el rendimiento de alumnos de educación especial basado en un test estandarizado con aquél de un grupo típico de norma nacional de la que la mayoría de los alumnos de educación especial han sido excluidos. De la misma manera, el ranking de países en comparaciones de exámenes internacionales para poder sacar conclusiones sobre la calidad del sistema educativo de un país son engañosas cuando diversos tipos y porcentajes de alumnos se presentan al examen en los distintos países. Las calificaciones medias de una evaluación internacional se sacó del 75% de los alumnos con 17 años de edad en los Estados Unidos, pero sólo del 9% más alto de los alumnos de 17 años de edad en Alemania y del 45% más alto en Suecia.

En este apartado también hay que añadir algo sobre otro tipo de estándar relativo—la práctica de "puntuar en curva", según la cual los maestros deciden desde un principio que los mejores alumnos recibirán una A, que los que están en medio una B, y los que están por debajo de éstos una C o D sin más definición sobre qué rendimiento se espera para cada calificación. Este tipo de estándar relativo simplemente clasifica a los alumnos. El problema es que aunque Kenny y Leila saquen una nota más alta que los demás y reciban una A, es posible que no hayan aprendido lo suficiente del contenido o que no sean capaces de rendir lo suficientemente bien para merecerse una A según un estándar absoluto de calidad de rendimiento. De igual manera, si el maestro y los materiales son suficientemente buenos, la clase entera podría ser capaz de trabajar muy bien y merecer una A. Lo importante es que mientras los estándares relativos tienen un lugar propio, el valor de los estándares absolutos se pasa muchas veces por alto. Al decir a los alumnos que estamos puntuando en curva, se les hace pensar que basta tener un mejor

rendimiento que otro, y que aquellos que están en el tercio inferior son mediocres, sin tener en cuenta sus esfuerzos y los nuestros, o que los estándares absolutos de lo que constituye un trabajo aceptable o excelente no son importantes.

Aplicar estándares forma parte del proceso inconsciente que utilizamos para hacer valoraciones. Tanto los estándares absolutos como los relativos representan métodos útiles para determinar el nivel de rendimiento de los alumnos. De hecho, los estándares absolutos muchas veces incorporan información relativa. ¿Cómo sabemos que los alumnos tienen que tener un 80% del examen de prácticas de laboratorio correcto para destacar en química? Porque a partir de nuestra experiencia, hemos comprobado que los mejores alumnos han sacado al menos un 80% en el examen de prácticas de laboratorio de ciencias. En la mayoría de los casos, contestaremos a la pregunta de "cómo vamos" consultando tanto los estándares absolutos como los apropiados grupos de referencia.

La utilización de resultados de exámenes para la toma de decisiones

Una vez que se ha decidido comparar el rendimiento de alumnos con estándares absolutos o relativos, podemos optar entre varias técnicas para resumir los resultados de la evaluación. A la vez que utilizamos estos métodos para resumir, hay que tener en cuenta que hay aún mucho más sobre el rendimiento del alumno que no nos revela la calificación. Cualquier forma de resumir produce un equilibrio entre la brevedad y la descripción detallada. A nuestro juicio la información descriptiva proporcionada por las evaluaciones alternativas constituye uno de sus atributos más atractivos. Sin embargo, habrá ocasiones cuando necesitaremos comunicar numéricamente los resultados. Hay tres formas básicas de presentar los números. Se puede presentar como una distribución de notas; dando la nota media, el medio o el modo; o mostrando el porcentaje de alumnos que ha alcanzado algún estándar absoluto.

La forma que utilizamos para resumir depende de los tipos de comparaciones que queremos hacer y si los criterios de calificación incluyen sólo una dimensión (escala) o varias.

Resumir una única dimensión

Consideremos en primer lugar a un caso simple, un sistema de calificación integral o de dimensión única.

La distribución de notas

Para ver la gama del rendimiento de los alumnos en una única dimensión, es necesario calcular simplemente cuántos alumnos recibieron cada nota posible. Se puede incluso dibujar la distribución de notas, utilizando ya sea el número o el porcentaje en bruto de alumnos que obtiene cada nota. Una representación del rendimiento de una clase, como por ejemplo la del cuadro 7.2, nos muestra si la mayoría de los alumnos sacan notas altas, bajas o notas medias. Esto puede ser de gran ayuda cuando no tenemos una idea preconcebida de cómo van a rendir los alumnos. Se puede utilizar tales gráficos para la monitorización de cómo lo hacemos con alumnos de un año a otro. Los investigadores llaman a la medida inicial "información de fondo".

El cuadro 7.3 ilustra la distribución del rendimiento (en una escala) en dos temas distintos de historia que CRESST ha utilizado en su investigación. Obsérvese que el gráfico muestra que hubo más alumnos que sacaron notas altas (de 3.5 a 5) en el tema de la inmigración que en el de Lincoln-Douglas. ¿Qué podría sugerirnos este tipo de información sobre la fuerza relativa de enseñanza en estas dos áreas?

Nota media. Otra forma de averiguar cómo van los alumnos es calcular escalas numéricas resumidas del rendimiento de la clase utilizando el promedio (la media aritmética), el medio (la mitad que está por encima y la que está por debajo), o el modo (la nota más frecuente). Estas escalas nos muestran cómo va la mayoría de los alumnos. Constituyen un código útil para comunicarse con otros.

Si un colega nos pregunta qué tal van los alumnos con las ecuaciones de oxidación-reducción, podemos recurrir a la estadística para contestarle "En una escala de 8 puntos, la media es de 6,8". Nuestro colega podrá entonces apreciar dónde se agrupa la mayoría de los alumnos y comparar este rendimiento con su propia clase, con los alumnos del año pasado o con lo que entiende sea capaz de hacer un alumno que se encuentra en la media de "6.8".

Nivel y porcentaje. Si utilizamos un estándar absoluto podemos decidir qué calificación representa el dominio o quizás podemos utilizar un estándar doble de rendimiento satisfactorio y ejemplar. Por ejemplo, en una escala de 5 puntos, un 3 puede representar el dominio en el primer sistema. En el sistema doble, un 3 puede representar un rendimiento satisfactorio y un 4 o más puede ser necesario para alcanzar el nivel de rendimiento ejemplar. Por consiguiente, podemos encontrar que un 10% de nuestros alumnos lograron un 4 o más alcanzando el nivel de rendimiento ejemplar y que otro 50% de los alumnos obtuvo un 3, y alcanzó el nivel de rendimiento satisfactorio. Esto se puede representar en un diagrama pastel para ilustrar qué proporción de alumnos pertenece a cada categoría (véase cuadro

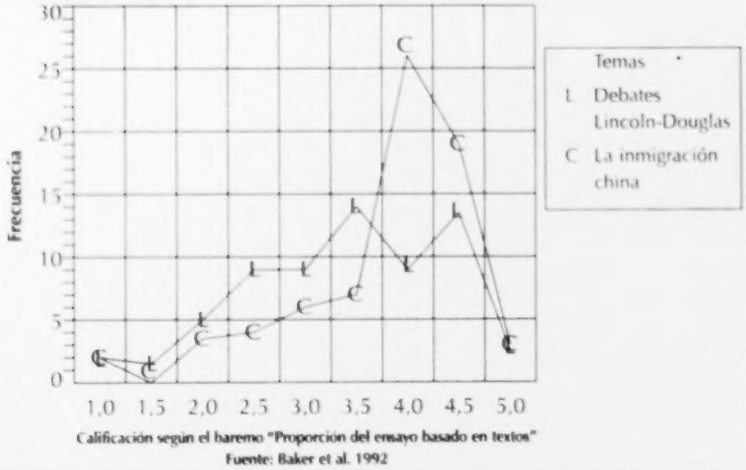
Cuadro 7.2
Distribución de notas de alumnos
Evaluación de procedimientos prácticos de biología

	(1)	(0)	(1)	(1)	(3)	(5)	(6)	(5)
							X	X
Número de alumnos						X	X	
						X	X	X
					X	X	X	X
					X	X	X	X
	X		X	X	X	X	X	X
	1	2	3	4	5	6	7	8

Notas posibles

X = Cada alumno que obtiene una calificación entre 1 y 8

Cuadro 7.3
Distribución de las calificaciones de alumnos en ensayos sobre dos temas de historia

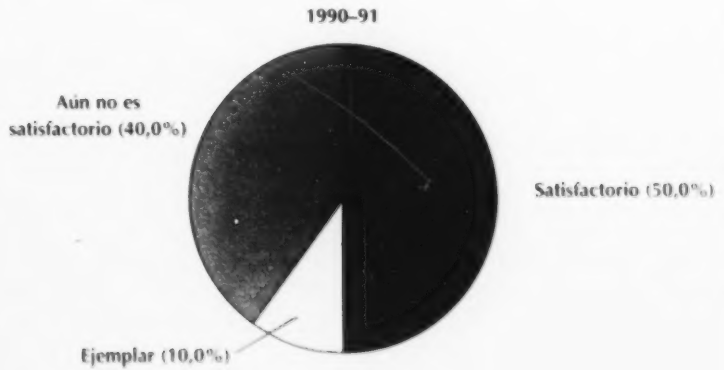


Calificación según el baremo "Proporción del ensayo basado en textos"
Fuente: Baker et al. 1992

7.4). Al igual que con las notas medias, los datos de dominio en porcentajes de un año o grupo se pueden comparar con aquellos de otro año o grupo.

Tendencias con el paso del tiempo. Sin tomar en cuenta si se utilizan distribuciones, notas medias o porcentajes de alumnos que alcanzan un nivel de rendimiento, quizás se quiera seguir la tendencia del rendimiento del alumno con el paso del tiempo. Uno se puede preguntar "¿Recibió la misma proporción de notas altas la clase de este año que la del año pasado?" "¿La nota media de este año fue superior o inferior de la del año pasado?" "¿Qué proporción de alumnos del último curso alcanzó este año el nivel de rendimiento ejemplar en comparación con la del año pasado?" Para un determinado alumno nos podríamos preguntar, "¿Qué tal es la nota de Justin en esta redacción persuasiva comparada con las notas que sacó en septiembre, noviembre y febrero?" Estas comparaciones longitudinales ayudan a poner en perspectiva el rendimiento de los actuales alumnos.

Cuadro 7.4
Porcentaje de alumnos que alcanzan estándares de rendimiento



Resumir varias dimensiones

Si se tienen varias dimensiones de rendimiento para resumir, hay dos opciones: (1) se puede sumar todas las notas o sacar la media—ambos métodos dan la misma visión global de lo que se hizo, o (2) se puede presentar por separado gráficas, medias o porcentajes para cada dimensión.

Si se suman las notas o se saca la media, quizá se quiera dar más peso a unas dimensiones que a otras en el caso de que sean más importantes para sus objetivos didácticos. Por ejemplo, aunque se califique la expresión escrita según las convenciones gramaticales, el estilo y la coherencia, se puede decidir dar más peso a la dimensión de coherencia—por ejemplo, multiplicando estas calificaciones por 1,5 ó 2—en comparación con la gramática y el estilo al presentar un resumen global del trabajo del alumno.

Hay un cierto "toma y daca" cuando sacamos la media o sumamos los criterios multidimensionales. Mientras nos hacemos una idea general del rendimiento, tenemos que estar conscientes de que las notas medias pueden esconder tipos de rendimiento muy diferentes. Por ejemplo, unos alumnos con una nota media de 7 pueden tener destrezas muy buenas de representación de problemas pero destrezas muy deficientes de resolución de problemas, mientras que otros alumnos pueden sacar un 7 en todas las dimensiones. Si hace falta ver tales distinciones en los resultados para así tomar decisiones didácticas, quizá se quiera presentar por separado los resultados para cada dimensión o para determinados baremos claves.

También nos podemos preguntar "¿qué tal vamos?" con respecto a cada dimensión. Por ejemplo, en mi tarea de evaluación de matemáticas me puedo preguntar ¿qué tal van mis alumnos en la comunicación, en la aplicación de conceptos matemáticos o en la utilización de fórmulas? Una estrategia útil para tratar los objetivos multidimensionales es la de observar la proporción de sub-baremos cuando el rendimiento fue suficiente o más. En nuestros ejemplos de tres sub-baremos podremos resumir nuestros resultados averiguando qué porcentaje de alumnos recibió una calificación de suficiente o más en una dimensión, en dos y en las tres. El cuadro 7.5 proporciona un ejemplo de esta estrategia.

Muestras de trabajos de alumnos

Sin tener en cuenta el método que se elige para presentar las conclusiones—bien de una sola dimensión, la media de varias dimensiones, o como una recopilación de varias dimensiones distintas—y si se presentan o no las tendencias con el transcurso del tiempo, las muestras de trabajos de alumnos ayudan a ilustrar los resultados y a tomar decisiones. Los números en sí no nos dicen todo lo que necesitamos saber. No queremos reducir todo a números y así perder la riqueza de las respuestas de los alumnos. Aún más importante, no queremos perder de vista la calidad del rendimiento del alumno y lo que significa un trabajo de calidad.

Cuadro 7.5
Resumir los criterios multidimensionales

Porcentaje de alumnos calificados con "suficiente o más" en un sub-baremo	Porcentaje calificado con "suficiente o más" en dos sub-baremos	Porcentaje calificado con "suficiente o más" en tres sub-baremos
100%	67%	35%

Al considerar la pregunta "¿qué tal vamos?" se podría seleccionar muestras de rendimiento que representen los mejores, los normales y los más deficientes niveles de rendimiento. Estos modelos comunican claramente a otros maestros, y muchas veces a los padres, la gama de rendimiento y dónde quedarían matriculados determinados alumnos. Si se archivan los mejores exámenes o incluso modelos de exámenes deficientes, normales, y extraordinarios, se puede observar cómo progresa el nivel de rendimiento general para cada grupo. ¿El informe excelente de prácticas de laboratorio que se hizo hace cinco años nos parece hoy sólo regular? Si la respuesta es afirmativa, quiere decir que estamos haciendo bien nuestro trabajo. ¿El periódico que se hizo en grupo sobre "La vida de los romanos" de años anteriores cuya calidad fue normal nos parece hoy excepcional al compararlo con los productos de hoy? Si es así podemos concluir que nos queda trabajo por hacer. Las muestras de rendimiento pueden cumplir el mismo propósito que los resúmenes numéricos cuando se toman decisiones informales para la clase.

¿Cómo podemos hacerlo mejor?

En nuestra opinión el primer objetivo de la evaluación es el de proporcionar retroalimentación para mejorar los logros de alumno individualmente, la didáctica del aula y los programas educativos. Si después de investigar resultados individuales y de grupo, averiguamos que no cumplimos los objetivos previstos, necesitamos identificar unas estrategias para poder mejorar. La evaluación diagnóstica identifica los tipos de cambios que se necesitan si esperamos mejorar teniendo en cuenta tanto los modelos como los procesos de rendimiento.

Entender los procedimientos del alumno

Para que las evaluaciones alternativas puedan responder a la pregunta de "¿cómo podemos mejorar?", debemos incluir en nuestras tareas y criterios posibilidades de observación y documentación de los procedimientos de los alumnos así como sus resultados. Si queremos saber cómo ayudar a los alumnos a hacer mejores presentaciones en grupo, necesitamos resultados referentes a cómo fueron planeadas las presentaciones, cómo fueron asignados los roles y cómo colaboraron los alumnos para realizar la tarea. La clave para establecer un diagnóstico es entender las causas o precursores del rendimiento. Aunque nunca podremos estar completamente seguros de qué tipo de didáctica produce qué resultados, necesitamos tomar en consideración algunas conjeturas con base, o mejor dicho, con hipótesis en cómo se construye un rendimiento bueno o excelente. Para hacer esto, necesitamos saber cómo se produce un rendimiento específico.

Muchas veces se recopila información diagnóstica al margen de la evaluación de los resultados. La fuente más rápida y rica de información de procesos es la de observar a los alumnos mientras llevan a cabo una tarea y, en momentos apropiados, interrumpirles individualmente de vez en cuando para preguntarles: ¿Qué has hecho para llegar a este punto? ¿Por qué hiciste aquello? ¿Qué podrías hacer ahora? Podemos incluso pedir a los alumnos que escriban en diarios sus reflexiones sobre su trabajo a lo largo del proceso; o quizás podemos caminar entre los alumnos mientras trabajan y apuntar notas breves de cara al futuro. Otras veces podemos tener con los alumnos sesiones en las que nos informen de su actividades y procesos para luego resumir los resultados en nuestros archivos anecdóticos.

A nivel centro, se puede llevar a cabo la monitorización de los procedimientos de los alumnos de varias formas: (1) observaciones formales en el aula, (2) grabación en vídeo, (3) transcripción, (4) comentarios de los propios compañeros, (5) charlas profesor-alumno, o incluso (6) análisis de documentos, un procedimiento para la recopilación y análisis de elementos fundamentales del aula—programas, evaluaciones, muestras de planificaciones de clases, muestras de trabajos seleccionados de alumnos, y carpetas de maestros o alumnos.

Podemos analizar esta información de procesos buscando modelos de conducta y relacionarlos con los resultados. ¿Los buenos alumnos se dirigían a la tarea de forma significativamente diferente a los alumnos no tan buenos? ¿Qué clase de ideas equivocadas tenían los que lo hicieron peor y cómo podrían estar relacionadas con un profundo malentendido de lo que se enseñaba? ¿Qué tipo de errores cometieron los alumnos menos aptos? ¿En qué parte del proceso de realización de la tarea tuvieron dificultades los alumnos? Esta continua retroalimentación sobre cómo están realizando una tarea los alumnos proporciona una valiosa información sobre cómo podemos ayudar a los alumnos a mejorar.

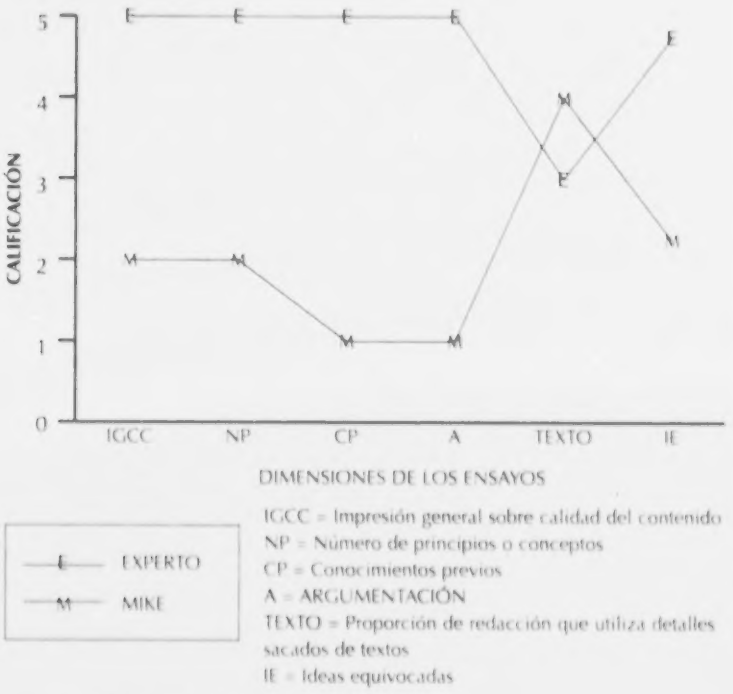
Perfiles de rendimiento

Si se está utilizando resultados de evaluaciones formales para fines diagnósticos, éstos deben poseer dos características: (1) un perfil, baremo, o conjunto de criterios que describe los aspectos constituyentes y de procesos del rendimiento y (2) razones válidas o marco teórico que sostiene la relación entre los componentes o procesos de la tarea y los resultados. Cuando se tienen criterios de la tarea basados en principios con una buena base teórica, se puede examinar los perfiles de rendimiento del alumno para identificar áreas de puntos fuertes y débiles relativos—para individuos, grupos, la clase entera, la escuela, etcétera. Por ejemplo, el cuadro 7.6 ilustra los puntos fuertes y débiles de la redacción de historia de Mike sobre el debate Lincoln-Douglas por medio de la representación en gráfica de sus calificaciones en seis dimensiones junto con el rendimiento teórico de un experto en historia, proporcionado por una investigación previa en CRESST (Baker et al. 1992). El cuadro 7.6 sugiere que comparado con el experto en historia, Mike incluyó pocos conocimientos previos globales y pocos principios históricos en su redacción, mostró una dependencia demasiado fuerte en un texto recientemente leído, construyó un argumento relativamente pobre, y utilizó varias ideas equivocadas.

Al utilizar la evaluación para fines diagnósticos, se querrá tener en cuenta la relación entre los sub-baremos de rendimiento y la calidad global del rendimiento. Su papel como diagnóstico se parece a aquel de un científico de la conducta; se está generando suposiciones comprobables sobre causa y efecto. ¿Cuál es la diferencia entre los perfiles de alumnos que rinden bien y los que no? ¿Qué dimensiones del rendimiento parecen ser las más importantes si queremos que los alumnos mejoren? ¿Cómo están relacionadas las distintas dimensiones? ¿Cuál debe enseñarse en primer lugar? Por ejemplo, si los alumnos que siempre argumentan de forma excelente tienen perfiles que son igualmente altos en conceptos como "la referencia a la información actual", "la utilización de hechos reales", y "la utilización del humor", entonces se querría consultar los perfiles de los alumnos con más dificultades para ver en cuál de estas dimensiones fallaban más. Si se averigua que los que no saben argumentar utilizan el humor y se refieren a hechos reales en sus argumentos pero fallan en la utilización de hechos de apoyo, se podría empezar a mejorar sus rendimientos trabajando con esta destreza.

A un nivel centro o distrito, cuando queremos fortalecer la enseñanza, nuestro enfoque está en el rendimiento de grupo en lugar de en el individual. Al examinar los resultados de grupo, es necesario observar tanto a los subgrupos como el rendimiento de sub-baremos. Los resúmenes a nivel de clase y centro muchas veces esconden muchos tipos de conocimientos previos y de experiencias de subgrupos identificables, como por ejemplo chicos, chicas, alumnos nuevos en el centro, hablantes no naturales de inglés, alumnos matriculados en determinados cursos, etcétera. Por ejemplo, el cuadro 7.7 ilustra perfiles de rendimiento separado entre chicos y chicas en una redacción de historia. El rendimiento está representado en seis dimensiones, y se puede ver que las chicas tuvieron calificaciones más altas que los chicos en todos los puntos del baremo, aunque la diferencia es mayor en unas dimensiones que en otras. Si damos por sentado que descartamos las

Cuadro 7.6
Perfiles de calificación de un experto y de un alumno en ensayos de historia



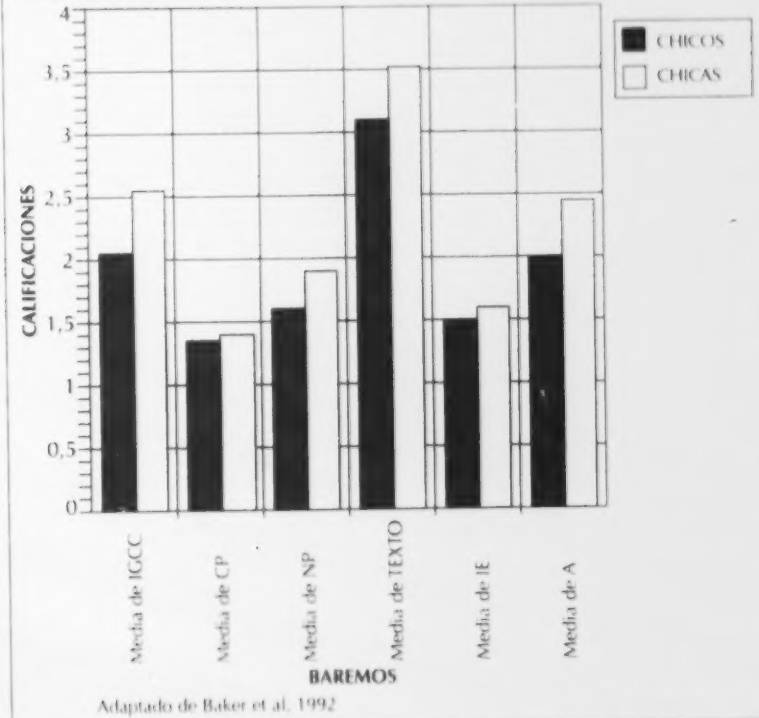
IGCC = Impresión general sobre calidad del contenido
NP = Número de principios o conceptos
CP = Conocimientos previos
A = ARGUMENTACIÓN
TEXTO = Proporción de redacción que utiliza detalles sacados de textos
IE = Ideas equivocadas

Adaptado de Baker et al. 1992

diferencias de imparcialidad de los calificadores, ¿qué podría significar estas diferencias de subgrupos en la toma de decisiones pedagógicas?

Si se desea llegar a todos los alumnos, se querrá saber si algunos subgrupos de alumnos tienen perfiles distintos a otros. Por ejemplo, ¿los chicos y chicas que sacan notas altas en la resolución de problemas matemáticos lo consiguen de la misma manera? O, entre el grupo de los que redactan de forma insatisfactoria ¿los alumnos nuevos tienen necesidades didácticas distintas a los que llevan tres o más años en el centro? Entre los "que apenas reprueban" y los "que apenas aprueban", ¿encontramos perfiles de rendimiento parecidos o diferentes? ¿Se parecen estos grupos fronterizos en sus calificaciones en gramática y utilización del lenguaje?

Cuadro 7.7
Perfiles de rendimiento entre chicos y chicas en ensayos de historia



Adaptado de Baker et al. 1992

¿Hay una dimensión de rendimiento que separa estos grupos fronterizos, como por ejemplo "organización", que podría dar lugar a enfocar la enseñanza de atención especial? Lo importante aquí es que cuando se observa a los resultados de grupo, no siempre se obtienen buenas pautas diagnósticas. No sólo hay que saber en qué áreas los alumnos de bajo rendimiento necesitan más atención, sino también hay que saber quiénes son estos alumnos.

Al igual que la nota media de la clase quizá no revele el hecho de que dos o tres alumnos no fueran capaces de llevar a cabo la tarea, los resúmenes de grupo pueden dar la falsa impresión de que todos los alumnos están rindiendo más o menos al mismo nivel. Una parte de la misión diagnóstica es la de averiguar qué alumnos o grupos no se reflejan suficientemente bien en el resumen para proporcionar resúmenes adecuados de sus rendimientos.

La utilización de sistemas de evaluación: Carpetas de trabajo como ejemplo modelo

Dadas las limitaciones cuando se utiliza una sola tarea de evaluación o un examen para generalizar acerca de un alumno, clase o centro en particular, sugerimos que se utilice varias tareas o situaciones para recopilar información sobre un alumno antes de tomar decisiones trascendentales. Un enfoque longitudinal a la evaluación pone en perspectiva los resultados de cualquier evaluación. A su vez, múltiples mediciones de los mismos resultados proporcionan opiniones alternativas de rendimiento que se combinan para crear una visión más completa del logro del alumno.

Muchos maestros han recurrido a la evaluación basada en carpetas de trabajo como estrategia para la creación de un sistema de evaluación de clase que incluye múltiples mediciones efectuadas a lo largo del tiempo. Las carpetas de trabajo tienen la ventaja de contener varias muestras de trabajo de un alumno ordenadas de manera deliberada. Las carpetas bien concebidas incluyen muestras que representan tanto trabajos en curso como muestras "modelo", reflexiones del alumno sobre su trabajo y los criterios de la evaluación. Arter y Spandel (1992) resumen los tipos de preguntas que los maestros debieran hacerse al utilizar carpetas de trabajo u otros sistemas de evaluación de conjunto:

1. ¿Hasta qué punto es representativo el trabajo incluido en la carpeta con respecto a lo que puede realmente hacer el alumno?
2. ¿Representan las muestras de la carpeta un trabajo dirigido? ¿Trabajo independiente? ¿Trabajo en grupo? ¿Se informa sobre la ayuda que recibieron los alumnos?
3. ¿Los criterios de evaluación para cada muestra y para la carpeta en conjunto representan las dimensiones más relevantes o útiles del trabajo del alumno?
4. ¿Hasta que punto corresponden a fines didácticos importantes o a tareas auténticas las muestras de la carpeta?
5. ¿Requieren las tareas o partes de ellas habilidades externas?
6. ¿Existe algún método para asegurar que las carpetas de trabajo se revisen de forma constante y que los criterios se apliquen con precisión?

La utilización de exámenes: El primer y último paso de la evaluación alternativa

A lo largo de este capítulo hemos hablado de la utilización de exámenes como si fueran el producto final del ciclo de desarrollo. Sin embargo es evidente que a menos que se considere la utilización de exámenes antes de la compra o diseño de una evaluación, es casi imposible conseguir la información que realmente se necesita. La evaluación, al igual que la enseñanza, requiere la consideración simultánea de muchos temas.

En este libro hemos planteado los temas conceptuales más importantes, si no todos los técnicos, en la evaluación alternativa. Nuestra lista es larga pero en absoluto exhaustiva. El campo de la evaluación alternativa está evolucionando con tanta velocidad que los cánones de hoy son las advertencias del mañana.

La creación y utilización de evaluaciones de rendimiento eficaces puede ser complicada. Si éste es su primer acercamiento, intente absorber primero las ideas más importantes. Sus evaluaciones probablemente mejorarán y con el tiempo los detalles se volverán más asequibles mientras uno se acostumbra a los conceptos y a la terminología. Al tratarse de un proceso iterativo, se plantearán cuestiones una y otra vez, cada vez con mayor experiencia y comprensión.

Esperamos que este manual les ayude a abrirse camino entre los matorrales de la siempre en aumento información sobre la evaluación alternativa para que puedan encontrar un sendero abierto hacia una evaluación más didácticamente sensible, fuerte, equitativa, y útil.

Referencias bibliográficas

Arter, J., y V. Spandel. (Primavera 1992). "Using Portfolios of Student Work in Instruction and Assessment." *Educational Measurement: Issues and Practice* 11, 1: 36-44.

Baker, E.L., P.R. Aschbacher, D. Niemi, y E. Sato. (1992). *CRESST Performance Assessment Models: Assessing Content Area Explanations*. Los Angeles: University of California, Center for Research on Evaluation, Standards and Student Testing.

Braddock, R., R. Lloyd-Jones, y L. Shoer. (1963). *Research in Written Composition*. Champaign, Ill.: National Council of Teachers of English.

Committee to Develop Standards for Educational and Psychological Evaluation. (1985). *Standards for Educational and Psychological Tests*. Washington, D.C.: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Elley, W.B., I.H. Barham, H. Lamb y M. Wyllie. (1976). "The Role of Grammar in a Secondary School English Curriculum." *Research in the Teaching of English* 10, 1: 5-21.

Herman, J.L. (1991). "Research in Cognition and Learning: Implications for Achievement Testing Practice." En *Testing and Cognition* (págs. 154-165), editado por M.C. Wittrock y E.L. Baker. Englewood Cliffs, N.J.: Prentice Hall.

Herman J.L., E.L. Baker, M. Gearhart, y A. Whittaker. (1991). "Stevens Creek Portfolio Project: Writing Assessment in the Technology Classroom." *Portfolio News* 2, 3: 7-9.

Shepard, L.A., y K. Cutts-Dougherty. (Abril 1991). "Effects of High-Stakes Testing on Instruction." Ponencia presentada en la reunión anual de la American Educational Research en Chicago, Ill.

123

Acerca de los autores

Joan L. Herman es Associate Director, Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles, Graduate School of Education, 10920 Wilshire Boulevard, Suite 900, Los Angeles, CA 90024.

Pamela R. Aschbacher es Assistant Director, Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles, Graduate School of Education, 10920 Wilshire Boulevard, Suite 900, Los Angeles, CA 90024.

Lynn Winters es Assistant Superintendent, Research, Planning, Evaluation, Long Beach Unified School District, Administration Building, 1515 Hughes Way, Long Beach, CA 90810.

Guía Práctica para una Evaluación Alternativa

En *Guía Práctica para una Evaluación Alternativa*, Joan Herman, Pamela Aschbacher y Lynn Winters nos ofrecen consejos convincentes para poder crear y utilizar métodos alternativos que midan los logros del alumno. Nos presentan un modelo que une la evaluación con el currículo y la docencia basado en las últimas teorías de aprendizaje y cognición.

Herman, Aschbacher y Winters repasan los fines de la evaluación y dan un argumento sustancial base de sus estrategias alternativas. La esencia del libro es la iluminación de varios temas claves relacionados con la evaluación que reafirma nuestra convicción de que hay que proporcionar a las tareas de evaluación los elementos más importantes de la práctica docente.

Association for Supervision and Curriculum Development
Alexandria, Virginia, USA

CRESST

National Center for Research on Evaluation, Standards, and Student Testing
University of California, Los Angeles
Los Angeles, California, USA

$15.
ISBN: 0-87120-28!